Populatie synthese

Een verkenning van methoden en tools voor SIVMO

Panteia

Populatie synthese

Een verkenning van methoden en tools voor SIVMO

Auteur(s)
Jan Kiel

Opdrachtgever(s)
SIVMO

Gepubliceerd
Zoetermeer, 2-2-2026

Projectnummer
11342

Versie
1.0

Status
Definitief

De verantwoordelijkheid voor de inhoud berust bij Panteia. Het gebruik van cijfers en/of teksten als toelichting of ondersteuning in artikelen, scripties en boeken is toegestaan mits de bron duidelijk wordt vermeld. Vermenigvuldigen en/of openbaarmaking in welke vorm ook, alsmede opslag in een retrieval system, is uitsluitend toegestaan na schriftelijke toestemming van Panteia. Panteia aanvaardt geen aansprakelijkheid voor drukfouten en/of andere onvolkomenheden.

Inhoudsopgave

	Samenvatting	**5**
1	**Inleiding**	**8**
1.1	Achtergrond	8
1.2	Doel	8
1.3	Aanpak	9
1.4	Leeswijzer	10
2	**Methoden populatie synthese**	**12**
2.1	Definitie en rol van populatie syntheses	12
2.2	Indeling in methodologische families	14
2.3	Beschrijving van families en methoden	15
2.3.1	Iteratieve ophoogmethoden	15
2.3.2	Optimalisatie-gebaseerde methoden	16
2.3.3	Probabilistische reconstructie	18
2.3.4	Simulatieve of generatieve modellen	20
2.3.5	Datafusie en hybride methoden	22
2.4	Samenvattende vergelijking van onderzochte methoden	24
3	**Tools en software**	**27**
3.1	Typologie en methode	27
3.2	Samenvattende vergelijkingstabel	28
3.3	Toolbeschrijvingen	29
3.3.1	SigPopu (Nederland)	29
3.3.2	Quad (Nederland)	30
3.3.3	Octavius (Nederland)	31
3.3.4	PopulationSim (VS)	33
3.3.5	PopGen (VS)	34
3.3.6	PopSynWin (VS)	35
3.3.7	MATSim population synthesizer (Switzerland)	37
3.3.8	SynthPop (VK)	38
3.3.9	Overige synthesizers	39
4	**Bevindingen uit interviews**	**42**
4.1	Overzicht van geïnterviewde partijen	42
4.2	PTV	43
4.3	Department for Transport (DfT, Verenigd Koninkrijk)	46
4.4	TNO	48
4.5	Centraal Bureau voor de Statistiek (CBS)	50
4.6	Goudappel	52
4.7	Significance	55

4.8	Overkoepelende bevindingen	56
5	**Beoordeling**	**61**
5.1	Beschouwing van methoden	61
5.2	Beoordelingskader voor de tools	62
5.3	Samenvattende vergelijking van de tools	63
5.4	Relevantie voor de Nederlandse context	64
5.5	Tussenconclusie	65
6	**Richting en aanbevelingen**	**68**
6.1	Uitgangspunten	68
6.2	Mogelijke routes voor implementatie	69
6.3	Aanbevolen richting	70
6.4	Aandachtspunten en vervolgstappen	72
	Annexes	
Bijlage 1	Referenties	75
Bijlage 2	Literatuur review	79
Bijlage 3	Interview verslagen	155
Bijlage 4	Lijst met afkortingen	183
Bijlage 5	Begrippenlijst	185

Samenvatting

Populatiesynthese wordt steeds belangrijker voor verkeers- en vervoersmodellen in Nederland, zowel nationaal als regionaal/lokaal. Tegelijk verschuift de modelpraktijk van trip-based naar tour-based en (op termijn) activity- en agent-based modellen, waarin persoons- en huishoudkenmerken een grotere rol spelen. Dit vraagt om een populatie-invoer die reproduceerbaar is, schaalbaar kan worden toegepast en expliciet rekening houdt met Nederlandse databronnen, zoneringen en privacy-/toegangskaders.

Er bestaat geen methode die in alle situaties “het beste” is. Iteratieve ophoog-methoden zoals IPF en IPU zijn relatief eenvoudig en transparant, maar lopen sneller tegen grenzen aan bij hoge dimensionaliteit, lege cellen en complexe combinaties van marges. Probabilistische reconstructie en sampling bieden flexibiliteit en maken onzekerheid expliciet, maar vragen extra maatregelen om variatie tussen runs te beheersen en marges consistent te houden. Simulatieve en generatieve benaderingen zijn relevant voor toekomstgerichte toepassingen, maar zijn complexer en stellen zwaardere eisen aan aannames en validatie. Optimalisatie-gebaseerde methoden bieden in veel gevallen een robuuste basis omdat zij meerdere randvoorwaarden tegelijk kunnen hanteren en doorgaans stabiel en reproduceerbaar blijven, ook wanneer data onvolledig of ‘lastig’ is.

De inventarisatie laat zien dat de praktische bruikbaarheid in sterke mate wordt bepaald door randvoorwaarden buiten het algoritme: datakwaliteit, consistente definities van marges, beschikbaarheid van (micro)data, privacy- en toegangskaders, en het beheer van versies, workflows en validatie. In de Nederlandse context is bovendien expliciet onderscheid nodig tussen wat binnen een gecontroleerde omgeving kan/mag gebeuren en wat daarbuiten reproduceerbaar beschikbaar moet zijn. Daarom is een werkbare aanpak niet alleen een keuze voor een methode of tool, maar ook de inrichting van een standaardproces voor data-voorbereiding, margecontrole, uitvoering, (eventuele) integerisatie en validatie, inclusief eenduidige documentatie van aannames, variabelen, herkomst van marges en versies.

Voor de inrichting binnen SIVMO zijn drie implementatieroutes te onderscheiden: (A) standaardiseren van bestaande (veelal gesloten) tools, (B) overnemen en aanpassen van een bestaande open synthesizer, of (C) het ontwikkelen van een nieuwe, generieke synthesizer. Op basis van de beoordeling en interviews ligt Route B het meest voor de hand: een bestaande open synthesizer met een optimalisatiebasis als uitgangspunt nemen en gericht aanpassen aan de Nederlandse context. Daarmee worden toetsbaarheid en overdraagbaarheid gecombineerd met beheersbaarheid, mits governance expliciet wordt ingericht. Tegelijk vraagt SIVMO een expliciete keuze over de gewenste output (gewichten/steekproef versus volledig geïntegeriseerde populatie) én een uitvoeringsscenario: een basis binnen gecontroleerde omgeving, een basis buiten gecontroleerde omgeving, of een hybride opzet waarin een reproduceerbare basis buiten de gecontroleerde omgeving wordt gecombineerd met verrijking/doorrekening binnen een gecontroleerde omgeving voor detailtoepassingen.

Belangrijke vervolgstappen zijn het vastleggen van functionele eisen (variabelen, schaalniveaus en consistentie-eisen), het uitwerken van een referentieworkflow inclusief kwaliteits- en validatiestappen, het expliciteren van governance (eigenaarschap, versiebeheer, wijzigingsprocedure, documentatiestandaard) en het uitvoeren van een pilot op een beperkt studiegebied om datastromen, margeopbouw, uitvoeringstijd en validatie in de praktijk te testen. Zo ontstaat een robuuste basis voor bredere toepassing en latere standaardisering, zonder directe afhankelijkheid van gesloten implementaties of de risico's en kosten van volledige nieuwbouw.

1

1 Inleiding

1.1 Achtergrond

Bij de SIVMO partners is sprake van een groeiende behoefte aan meer inzicht in populatiesynthese op verschillende geografische schaalniveaus. Waar eerder vooral op nationaal en bovenregionaal niveau synthetische populaties werden gebruikt (in het LMS en NRM), groeit nu ook de vraag naar toepassing op regionaal en lokaal niveau. Dit sluit aan bij de verschuiving van trip-based naar tour-based en activity-based modellen, waarin gedragskenmerken van personen en huishoudens centraal staan. Voor dergelijke toepassingen is een stabiele en goed gespecificeerde populatie-invoer belangrijk.

Binnen SIVMO is afgesproken om te onderzoeken welke methoden en tools voor populatiesynthese in aanmerking komen voor bredere toepassing of standaardisering. Doel is niet om direct een definitieve keuze te maken, maar om inzichtelijk te maken wat beschikbaar is, wat goed functioneert, en wat in de praktijk minder toepasbaar is. Dit past binnen de bredere ambitie van SIVMO om het modelleringsinstrumentarium stapsgewijs te versterken.

Dit rapport vormt de eerste stap in een traject dat kan leiden tot een breed afgestemde implementatie van populatiesynthese binnen vervoers- en verkeers-modellen. Ook biedt het een fundament voor eventuele gezamenlijke keuzes rond standaardisering, hergebruik en verdere ontwikkeling van synthesetools.

1.2 Doel

Dit rapport heeft als doel een systematisch overzicht te geven van bestaande benaderingen voor populatiesynthese die relevant zijn voor de Nederlandse verkeers- en vervoersmodellen. Daarbij ligt de nadruk op de mate waarin deze methoden en tools bruikbaar, aanpasbaar en schaalbaar zijn voor toepassing in trip-based, tour-based en activity-based modellen zoals gebruikt in het SIVMO-netwerk.

Het onderzoek levert hiervoor drie hoofdbestanddelen op. Allereerst beschrijft en vergelijkt dit rapport de onderliggende methoden voor het synthetiseren van populaties aan de hand van een literatuurreview. Vervolgens inventariseren we de beschikbare tools die deze methoden in de praktijk gebruiken – zowel nationaal als internationaal, open en gesloten. Tot slot toetsen we de geschiktheid van deze methoden en tools aan de hand van praktijkervaring, interviews en literatuur.

De centrale onderzoeksvraag luidt:
Welke bestaande methoden en tools voor populatiesynthese zijn het meest geschikt voor gebruik binnen de Nederlandse modelpraktijk, en welke route is wenselijk voor standaardisering of verdere ontwikkeling binnen SIVMO?

Het eindresultaat van deze verkenning is een advies dat inzicht biedt in de mogelijke routes: het gebruiken van een bestaande oplossing, het aanpassen van een bestaande tool aan de Nederlandse context, of het (laten) ontwikkelen van een nieuwe, generieke populatiesynthesizer. De aanbevelingen zijn bedoeld om direct bruikbaar te zijn bij toekomstige modelaanpassingen of aanbestedingen binnen het SIVMO-netwerk.

1.3 Aanpak

De inventarisatie is uitgevoerd op basis van een combinatie van literatuuronderzoek, interviews en toetsing aan de praktijk. De aanpak is ontworpen om zowel de breedte van beschikbare methoden en tools in kaart te brengen als de diepgang te zoeken op aspecten die belangrijk zijn voor de Nederlandse modelpraktijk.

Het literatuuronderzoek bestond uit twee fasen. In de eerste fase is een overzicht gemaakt van methoden en technieken voor populatiesynthese aan de hand van wetenschappelijke artikelen, technische rapporten, projectdocumentatie en tool-documentatie. Hiervoor is gewerkt met expliciete zoekstrings (o.a. op kernbegrippen rond 'population synthesis' en relevante toepassingsdomeinen), uitgevoerd over een vastgelegde zoekperiode (in principe na 2010) en via meerdere bronnen (wetenschappelijke zoekmachines, publicatieplatformen en tool- en projectdocumentatie). Per zoekronde is vastgelegd welke bronnen uiteindelijk zijn meegenomen in de verdere analyse (zie literatuuroverzicht).

In de tweede fase is een selectie van bronnen verdiept, met aandacht voor de onderliggende aannames, rekenkundige principes en toepassingsmogelijkheden. Ook zijn concrete tools onderzocht en gecategoriseerd op basis van hun methodologische fundament, openheid, gebruik in de praktijk en geschiktheid voor doorontwikkeling.

Op basis van de literatuurverkenning zijn interviews gehouden met betrokkenen uit verschillende domeinen: modelontwikkelaars, beleidsmakers en onderzoekers. Deze gesprekken zijn gebruikt om literatuurbevindingen te toetsen en om aanvullende inzichten te verzamelen over datavereisten, validatie in de dagelijkse praktijk, beheer- en onderhoudsvragen, en wensen voor bruikbaarheid en transparantie.

De verzamelde inzichten zijn vervolgens vertaald naar een toetsingskader. Daarmee zijn de methoden en tools vergeleken op inhoudelijke, organisatorische en praktische criteria (zoals geschiktheid voor toepassing, schaalbaarheid, reproduceerbaarheid, robuustheid, validatie, beheer, onderhoud en toegankelijkheid en openheid). Dit resulteert in een onderbouwd overzicht van sterke en zwakkere punten, mogelijke toepassingen en beperkingen, en een eerste verkenning van kansrijke richtingen voor standaardisering of ontwikkeling binnen SIVMO.

Bij het opstellen van dit rapport is gebruik gemaakt van verschillende tools: ChatGPT, Elicit, NotebookLM, ResearchGate en Whisper. Deze tools zijn ingezet als ondersteuning bij het zoeken en ordenen van literatuur (bijv. het verbreden van zoektermen, het structureren van gevonden bronnen), het samenvatten van teksten en het vergelijken van kenmerken van methoden en tools. Whisper is gebruikt voor transcriptie/uitwerking van interviews. De uiteindelijke selectie van bronnen, interpretatie van resultaten en weging van bevindingen is door de auteurs uitgevoerd.

Omdat kenmerken zoals openbaarheid, licentie en onderhoudsstatus van tools snel kunnen veranderen, is een verificatiestap toegepast. Daarbij zijn toolkenmerken waar mogelijk gecontroleerd aan de hand van primaire bronnen zoals officiële documentatie, repository-informatie (bijv. licentiebestand en recente updates of releases) en projectwebsites, en waar relevant gespiegeld aan interviewinformatie. Onzekerheden en aannames zijn daarbij zo veel mogelijk expliciet gemaakt.

Noot: Dit onderzoek is met zorg uitgevoerd op basis van literatuur, documentatie en interviews met betrokken partijen. Toch blijft het mogelijk dat bepaalde beschrijvingen onvolledig zijn, dat interpretaties afwijken van de meest recente stand van zaken, of dat tools en methoden in de praktijk anders worden toegepast dan uit openbare bronnen blijkt. Populatiesyntheses ontwikkelen zich bovendien, waardoor keuzes en implementaties kunnen wijzigen in de periode tussen dataverzameling en afronding van dit rapport. Waar onzekerheden bestonden, zijn deze zo veel mogelijk expliciet gemaakt.

1.4 Leeswijzer

Het rapport is opgebouwd uit zes hoofdstukken en een aantal bijlagen:.

- **Hoofdstuk 2** beschrijft de methoden voor populatiesynthese. Het hoofdstuk start met een definitie en de rol van populatiesynthese, werkt de indeling in methodologische families uit en beschrijft vervolgens de families en methoden. Het hoofdstuk sluit af met een samenvattende vergelijking van de onderzochte methoden.
- **Hoofdstuk 3** behandelt de beschikbare tools en software voor populatiesynthese. Het hoofdstuk introduceert een typologie, geeft een samenvattende vergelijkingstabel en beschrijft daarna de belangrijkste tools afzonderlijk, aangevuld met een overzicht van overige synthesizers.
- **Hoofdstuk 4** bevat de bevindingen uit de interviews. Eerst wordt een overzicht van de geïnterviewde partijen gegeven, daarna volgen de bevindingen per organisatie en een afsluitende synthese van de overkoepelende inzichten.
- **Hoofdstuk 5** geeft de beoordeling. Het hoofdstuk beschouwt eerst de methoden, werkt vervolgens het beoordelingskader voor tools uit en presenteert een samenvattende vergelijking van de tools. Daarna wordt de relevantie voor de Nederlandse context besproken en volgt een tussenconclusie.
- **Hoofdstuk 6** beschrijft de richting en aanbevelingen. Het hoofdstuk formuleert uitgangspunten, werkt mogelijke routes voor implementatie uit, geeft een aanbevolen richting en benoemt aandachtspunten en vervolgstappen.
- De **bijlagen** bevatten de referenties, de uitwerking van de literatuurreview, de interviewverslagen, een lijst met afkortingen en een begrippenlijst.

2

2 Methoden populatie synthese

Dit hoofdstuk beschrijft de conceptuele en methodologische grondslagen van populatiesynthese. We starten met een definitie en typering van het gebruik in vervoersmodellen, gevolgd door een indeling van methoden in vijf families. Vervolgens behandelen we zestien afzonderlijke methoden in meer detail en sluiten we af met een samenvattende vergelijking.

2.1 Definitie en rol van populatie syntheses

In verkeers- en vervoersmodellen bedoelen we met populatiesynthese het systematisch opstellen van een realistische representatie van een populatie op basis van beschikbare gegevens. Dit gebeurt doorgaans door microdata (zoals persoons- of huishoudensenquêtes) te combineren met geaggregeerde randtotalen (zoals tabellen van het CBS), zodat een coherente en plausibele populatie ontstaat die als invoer kan dienen voor simulatiemodellen.

Een synthetische populatie bestaat uit afzonderlijke entiteiten – doorgaans personen en/of huishoudens – waaraan kenmerken zijn toegekend zoals leeftijd, huishoud-samenstelling, inkomen, opleidingsniveau of woonlocatie. Deze populatie moet enerzijds aansluiten op de geobserveerde verdelingen in de populatie (de marges of CBS-statistieken), en anderzijds voldoende gedetailleerd en intern consistent zijn om als basis te dienen voor gedragsmodellering, modaliteitskeuze, tijdstipkeuze of voertuigbezit (Harland et al., 2012; Ye et al., 2009).

Populatiesynthese is gericht op het genereren van een plausibel microbestand op basis van deels onvolledige en ongelijksoortige gegevens. Daarvoor kunnen uiteen-lopende technieken worden toegepast, van proportionele ophoogmethoden tot probabilistische reconstructies en scenarioafhankelijke simulaties. De keuze voor een bepaalde methode hangt onder meer af van het doel van de toepassing, de beschikbaarheid en kwaliteit van data, en het type vervoersmodel waarin de populatie wordt ingezet (Chapuis & Taillandier, 2019; Borysov et al., 2019).

Er zijn meerdere redenen waarom populatiesynthese noodzakelijk of wenselijk kan zijn:

1. *Gebrek aan volledige microdata.* Volledige gegevens over alle individuen in een populatie zijn zelden beschikbaar, vanwege privacybeperkingen, hoge kosten of wettelijke beperkingen. Synthese maakt het mogelijk om op basis van steek-proeven en geaggregeerde gegevens toch een volledige populatie te maken (Müller & Axhausen, 2010; Harland et al., 2012).
2. *Benodigd voor fine-grained modellen.* Moderne verkeersmodellen, zoals activity-based modellen (AcBM) of agent-based modellen (AgBM), vereisen informatie op individueel niveau. Synthese maakt het mogelijk om zulke data te genereren, met

voldoende detail voor analyses naar bijvoorbeeld inkomensverschillen of gezinsstructuren (Rich, 2018; Balać & Hörl, 2021).

3. *Integratie van meerdere databronnen.* Populatiesynthese maakt het mogelijk om verschillende databronnen – zoals enquêtes, statistieken en geodata – te combineren tot één samenhangend geheel (Ye et al., 2009; Beemster, 2016). Dit is vooral nuttig wanneer bepaalde variabelen niet gezamenlijk beschikbaar zijn in één bron.
4. *Scenarioverkenning en projectie.* Doordat populaties kunnen worden gesynthetiseerd op basis van alternatieve verdelingen of prognoses, kunnen ook toekomstscenario's worden gegenereerd. Dit is belangrijk voor uiteenlopende beleidsverkenningen (Müller, 2014; Rich, 2018).
5. *Privacybescherming.* Omdat synthetische populaties geen echte personen bevatten, maar statistisch vergelijkbare 'fictieve' agenten, zijn ze geschikt voor toepassingen waarin privacy belangrijk is. Dit maakt publieke gebruik eenvoudiger (Harland et al., 2012).

In dit rapport beperken we ons tot systematische, reproduceerbare methoden voor populatiesynthese. We focussen op methoden die geschikt zijn voor vervoersmodellen, en die dus rekening houden met kenmerken die mobiliteitsgedrag mede bepalen. In dit rapport onderscheiden we drie hoofdtypen van vervoersmodellen: trip-based modellen, tour-based modellen en activity-based modellen (AcBM). Elk modeltype stelt andere eisen aan de detaillering, structuur en stabiliteit van de populatie-invoer.

Trip-based modellen beschrijven verplaatsingen als afzonderlijke trips naar herkomst en bestemming. Deze modellen zijn relatief eenvoudig en vereisen minder gedetailleerde input op persoons- of huishoudniveau. Toch is ook hier een consistente populatie-invoer nodig, bijvoorbeeld voor het bepalen van trips naar leeftijdsgroepen, huishoudtypen of voertuigbezit.

Tour-based modellen gaan een stap verder door verplaatsingen te groeperen in samenhangende reeksen, bijvoorbeeld een woon-werkreis (huis–werk–huis). Dit vereist informatie over dagelijkse routines en gedragsstructuren. Kenmerken zoals arbeidsparticipatie, huishoudstructuur en tijdsbesteding moeten daarom correct en consistent in de populatie worden opgenomen.

Activity-based modellen beschrijven gedrag op het niveau van opeenvolgende activiteiten gedurende de dag. Ze vereisen een gedetailleerde en onderling consistente beschrijving van personen en huishoudens, inclusief interacties tussen gezinsleden (zoals gezamenlijke verplaatsingen of afstemming van activiteiten). Voor AcBM's is populatiesynthese onmisbaar: zonder stabiele en coherente populatie-invoer kunnen deze modellen niet functioneren. Vaak is de gesynthetiseerde populatie ook direct gekoppeld aan andere modelmodules, zoals voertuigbezit, activiteitengeneratie of vervoerwijzekeuze.

Voor elk van deze modeltypen geldt dat de eisen aan de populatie-invoer toenemen naarmate het model meer gedragsdetail en interne samenhang bevat. Waar bij trip-based modellen een beperkte segmentatie volstaat, vragen tour-based en vooral activity-based modellen om een expliciet gesynthetiseerde, realistische en stabiele populatie.

2.2 Indeling in methodologische families

Voor populatiesynthese bestaan uiteenlopende methoden, ontwikkeld binnen en voor domeinen zoals transport, demografie en epidemiologie. De technieken verschillen in opzet, databehoefte en rekenkundige aanpak. In dit rapport hanteren we een indeling in vijf methodologische families. Deze zijn geordend naar het dominante principe waarmee de synthetische populatie wordt gegenereerd.

Deze indeling is gebaseerd op overzichtsstudies van onder meer Rich et al. (2019), Chapuis en Taillandier (2019) en Müller en Axhausen (2010), en is aangevuld met recentere benaderingen zoals deep generative models en datafusietechnieken. Door methoden in families te groeperen, kunnen we ze systematisch beschrijven en vergelijken, los van specifieke softwaretools of nationale toepassingen.

De vijf onderscheiden families zijn:

1. *Iteratieve ophoogmethoden.* Methoden zoals Iterative Proportional Fitting (IPF) en Iterative Proportional Updating (IPU) passen microdata aan geaggregeerde randtotalen aan via herhaalde (iteratieve) proportionele correcties. Deze benadering is breed toegepast in verkeersmodellen en vormt vaak de basis voor basisjaarpopulaties (Choupani & Mamdoohi, 2016; Ye et al, 2009).
2. *Optimalisatie-gebaseerde methoden.* Deze methoden formuleren populatie-synthese als een optimalisatieprobleem, waarbij bijvoorbeeld afwijkingen tussen marges en gegenereerde populatie worden geminimaliseerd, of de entropie wordt gemaximaliseerd. Voorbeelden zijn Entropiemaximalisatie en Least Squares Matching. Ze worden vaak gebruikt in situaties waar precisie, interne consistentie of aanvullende randvoorwaarden belangrijk zijn (Müller, 2014).
3. *Probabilistische reconstructie.* Hierbij worden micro-eenheden getrokken uit steekproeven op basis van kansverdelingen die aansluiten op bekende marges. Dit kan met eenvoudige weging, of via complexere methoden zoals Monte Carlo sampling, Bayesiaanse netwerken of Combinatorial Optimisation Sampling (Borysov et al, 2019; Rahman & Fatmi, 2023).
4. *Simulatieve of generatieve modellen.* In deze benadering worden populaties gegenereerd op basis van gedragsregels of stochastische processen, zoals bij agent-based modellen. Hierbij ontstaan microbestanden als resultaat van simulatie, waarbij bijvoorbeeld huishoudvorming, migratie of opleidingsniveau wordt nagebootst. Voorbeelden zijn Rule-based populatiesimulatie of Agent-bases synthetic population generation (Chapuis & Taillandier, 2019; Hörl & Balać, 2020).
5. *Datafusie en hybride technieken.* Deze methoden combineren meerdere bronnen of technieken, bijvoorbeeld door record linkage, synthetische matching, data imputatie of machine learning. Ze zijn vooral relevant wanneer datasets onvolledig zijn of wanneer privacybeperkingen een rol spelen. Tot deze familie behoren ook nieuwere methoden gebaseerd op deep learning, zoals *variational autoencoders* (VAE), *generative adversarial networks* (GANs) en *diffusion models*. Deze worden gebruikt voor synthetische datageneratie, aanvulling van ontbrekende attributen of privacybescherming (Borysov et al., 2019; Albiston et al., 2024; Wu & Lyu, 2024b).

Elke familie omvat meerdere concrete methoden die gedeelde principes en toepassingslogica hebben. In de volgende paragraaf (2.3) worden de afzonderlijke methoden binnen deze families beschreven, met aandacht voor hun werking, toepassingsgebied, voordelen en beperkingen. In hoofdstuk 3 koppelen we deze methoden aan specifieke softwaretools.

2.3 Beschrijving van families en methoden

2.3.1 Iteratieve ophoogmethoden

Inleiding
De iteratieve ophoogmethoden vormen een klassieke en veelgebruikte benadering voor populatiesynthese, met een sterke positie in trip-based en tour-based modellen. Het principe is eenvoudig: een steekproef (bijvoorbeeld uit een huishoudens- of personenenquête) wordt aangepast zodat de marginale verdelingen van kenmerken overeenkomen met externe randtotalen, zoals CBS-statistieken.

Deze technieken zijn met name geschikt voor situaties waarin betrouwbare marges beschikbaar zijn, maar geen volledige populatiegegevens. De methode is deterministisch en goed reproduceerbaar, wat belangrijk is voor beleidsmodellen met transparante onderbouwingen.

Methoden
Binnen deze familie onderscheiden we de volgende technieken:

- *Iterative Proportional Fitting (IPF).* Dit is de bekendste en eenvoudigste vorm, waarbij celwaarden in een kruistabel herhaaldelijk worden aangepast om te voldoen aan opgelegde marges. De techniek is al decennia in gebruik en staat ook bekend onder namen als raking, Furness of Fratar bij 2D-toepassingen. Zie Choupani & Mamdoohi (2016).
- *Iterative Proportional Updating (IPU).* Deze uitbreiding van IPF (ontwikkeld door Ye et al, 2009) maakt het mogelijk om huishoud- en persoonskenmerken tegelijk te synthetiseren, met behoud van consistente structuur. Elk huishouden krijgt daarbij een gewicht dat in het iteratieproces wordt aangepast. IPU is met name nuttig bij multilevel-structuren. Toepassingen en varianten zijn onder andere beschreven door Müller & Axhausen (2010) en Harland et al. (2012).

Werking
De kern van deze methoden is het herhaaldelijk aanpassen van gewichten (IPU) of frequenties (IPF) totdat de marginale verdelingen overeenkomen met vooraf vastgestelde randtotalen. Meestal wordt een representatieve steekproef als 'seed' gebruikt. De aanpassingen vinden plaats per kenmerk (bijv. leeftijd, geslacht, huishoudtype), in opeenvolgende iteraties. Het proces stopt als alle marges binnen een vooraf bepaalde tolerantie zijn gebracht.

Toepassingen
Iteratieve methoden zijn breed toepasbaar en worden vaak gebruikt voor het opstellen van basispopulaties. Binnen Nederland maakt TNO gebruik van een synthesiser die is gebaseerd op IPF, maar dit is geen standaard beschikbare tool.

Voordelen

- Transparant en eenvoudig te begrijpen
- Reproduceerbare resultaten
- Lage rekenbelasting
- Weinig modelparameters nodig
- Ruim gedocumenteerd en ondersteund in software

Beperkingen

- Gevoelig voor lege celcombinaties of zeldzame categorieën
- Slechter schaalbaar bij hoge dimensionaliteit (>5 kenmerken)
- Geen expliciete modellering van correlaties tussen kenmerken
- Afhankelijk van kwaliteit en representativiteit van de steekproef

Reflectie op gebruik en validatie

Hoewel IPF en IPU een redelijke match kunnen geven met marginale verdelingen, ontbreekt de mogelijkheid om correlaties tussen kenmerken expliciet te modelleren. Bij microsimulaties, bijvoorbeeld als verplaatsingsgedrag samenhangt met huishoudstructuur, voertuigbezit en opleidingsniveau, is dat een belangrijk nadeel. Iteratieve methoden reconstrueren deze afhankelijkheden niet actief; ze zijn enkel aanwezig voor zover ze impliciet in de steekproef zitten.

Daarnaast blijkt uit praktijkervaring dat iteratieve methoden gevoelig zijn voor tekortkomingen in de inputdata. Bijvoorbeeld: als een bepaalde combinatie van kenmerken in de steekproef niet voorkomt, kan IPF die combinatie niet genereren, zelfs als deze wel plausibel is in de populatie. Hierdoor zijn deze methoden minder geschikt voor toekomstscenario's of contexten met structurele veranderingen.

Validatie van de gegenereerde populatie aan de hand van waargenomen data blijft belangrijk. In de LMS-backcast is gebleken dat het vergelijken van gesynthetiseerde kenmerken (zoals huishoudstructuren of voertuigbezit) met feitelijke waarnemingen veel inzicht geeft in de beperkingen van de methode en van de gebruikte steekproef.

2.3.2 Optimalisatie-gebaseerde methoden

Inleiding

Optimalisatie-gebaseerde methoden beschouwen populatiesynthese als een formeel optimalisatieprobleem. Het doel is om een synthetische populatie te vinden die enerzijds voldoet aan opgelegde randvoorwaarden (zoals CBS-statistieken) en anderzijds zo dicht mogelijk ligt bij een gewenste verdeling of zo min mogelijk afwijking introduceert. Hiervoor wordt een doelfunctie geminimaliseerd of gemaximaliseerd, zoals de Kullback-Leibler-divergentie, de som van kwadratische afwijkingen, of het verschil ten opzichte van een prior.

De uitkomsten van deze methoden zijn in principe stabiel, deterministisch en reproduceerbaar. Ze zijn goed inzetbaar in situaties met meerdere dimensies, aanvullende 'constraints' en een duidelijke scheiding tussen invoer en model-structuur. In tegenstelling tot iteratieve methoden worden alle marges en voor-waarden gelijktijdig meegenomen, wat de interne consistentie van de uitkomst versterkt.

Methoden

Binnen deze familie onderscheiden we de volgende technieken:

- *Entropiemaximalisatie.* Deze methode zoekt de verdeling met maximale entropie onder opgelegde randvoorwaarden. Dat betekent: genereer een verdeling die zo min mogelijk extra structuur aanbrengt in de populatie, tenzij dat nodig is om aan marges te voldoen. Dit principe is wiskundig verwant aan het minimaliseren van de Kullback-Leibler-divergentie ten opzichte van een uniforme verdeling. Entropie-maximalisatie wordt toegepast in bijvoorbeeld PopulationSim (Paul et al., 2017).
- *Least Squares Matching.* Hierbij wordt de som van kwadratische afwijkingen tussen gegenereerde marges en opgelegde marges geminimaliseerd. Deze aanpak is intuïtief, statistisch interpreteerbaar en geschikt voor extensies zoals gewichten of prioriteiten. Het nadeel is dat de gegenereerde verdeling niet per se voldoet aan alle marges als de 'constraints' zacht worden behandeld. Zie o.a. Müller (2014) voor een wiskundige onderbouwing en toepassingen in microsimulatie.
- *Kullback-Leibler-divergentie-minimalisatie.* In plaats van te zoeken naar een verdeling met maximale entropie, zoekt deze methode de verdeling die zo min mogelijk afwijkt van een bekende startverdeling (de prior), gegeven de rand-voorwaarden. Dit gebeurt door het minimaliseren van de zogenaamde Kullback-Leibler-divergentie tussen beide verdelingen. SigPopu volgt deze benadering, waarbij de prior doorgaans gebaseerd is op gewichten uit een steekproef of seed-bestand
- *Numerieke oplossers zoals Newton-Raphson*. Sommige tools, zoals Quad, gebruiken een Newton-Raphson methode om de optimale gewichten te vinden voor een synthetische populatie. Hoewel het model sterk lijkt op IPF (invoer: marges en microdata), is de rekentechniek anders: een systeem van evenwichts-vergelijkingen wordt numeriek opgelost. De aanpak is snel bij lage dimensies, maar kwetsbaar voor instabiliteit bij complexere situaties of lange tijdreeksen.

Werking

Bij deze methoden wordt een doelfunctie gedefinieerd (bijvoorbeeld: minimaliseer KL-divergentie of kwadratische afwijking), onder randvoorwaarden zoals marges, non-negativiteit of vaste groepsverdelingen. Afhankelijk van de implementatie worden oplossingsmethoden zoals Newton-Raphson, lineair programmeren of Lagrange-optimalisatie gebruikt. Het resultaat is een gewichtsverdeling of een gegenereerde populatie die het optimum van het systeem benadert.

Toepassingen

Optimalisatiemethoden zijn geschikt voor activity-based modellen, toekomst-scenario's en situaties met hoge dimensionaliteit. In Nederland wordt SigPopu gebruikt in combinatie met zonegewichten en een gewogen seedpopulatie. Quad gebruikt eveneens een optimalisatie-aanpak via Newton-Raphson. Internationaal worden PopulationSim en PopGen vaak genoemd als voorbeelden van tools met optimalisatiegebaseerde populatiesynthese.

Voordelen

- Levert stabiele, unieke oplossingen
- Robuust bij lege celcombinaties
- Uitbreidbaar met extra randvoorwaarden of beperkingen
- Goed reproduceerbaar en interpreteerbaar

Beperkingen

- Hogere rekenlast, vooral bij grote populaties of veel constraints
- Complexere implementatie dan IPF
- Vereist nauwkeurige randdata en parameterinstellingen
- Minder intuïtief dan proportionele benaderingen.

Reflectie op gebruik en validatie

Optimalisatie-gebaseerde methoden zijn aantrekkelijk vanwege hun transparantie, reproduceerbaarheid en het vermogen om meerdere randvoorwaarden tegelijk te verwerken. Ze leveren consistente uitkomsten en zijn in staat om populaties te genereren die structureel voldoen aan opgelegde marges. Toch is dit geen garantie voor een plausibele uitkomst: de effectiviteit hangt sterk af van de kwaliteit van de invoerdata, de gekozen 'constraints' en de aannames in de doelfunctie.

In de praktijk kunnen afwijkingen ontstaan als marges onvolledig zijn of de prior (startverdeling) onvoldoende representatief is. In sommige gevallen leidt dit tot onrealistische ophoogfactoren of het verdwijnen van bepaalde subgroepen. Dit is bijvoorbeeld waargenomen bij de toepassing van Quad in backcastanalyses, en bevestigt dat ook wiskundig correcte oplossingen inhoudelijk tekort kunnen schieten. SigPopu biedt hier meer flexibiliteit, maar vereist eveneens zorgvuldige setup en toetsing (Significance, 2024)

Om die reden is validatie van de gesynthetiseerde populatie belangrijk. Niet alleen moet worden gecontroleerd of marges worden gehaald, maar ook of de interne structuur en verdelingen plausibel zijn. Vergelijking met observaties, toetsing op stabiliteit en beoordeling van correlatiestructuren vormen daarbij noodzakelijke stappen voor betrouwbaar gebruik in verkeersmodellen (Harland et al., 2012; Choupani & Mamdoohi, 2016; Rich et al., 2019).

2.3.3 Probabilistische reconstructie

Inleiding

Probabilistische reconstructie omvat methoden die een synthetische populatie genereren door herhaalde trekkingen uit een steekproefbestand. De onderliggende veronderstelling is dat een populatie niet uniek is, maar gerepresenteerd kan worden door meerdere plausibele realisaties, zolang de uitkomsten statistisch in overeenstemming zijn met bekende marges. In plaats van deterministisch toe te werken naar een enkele oplossing, staat bij deze methoden het expliciet modelleren van variatie centraal.

Methoden

- *Monte Carlo sampling.* De eenvoudigste vorm is het trekken van huishoudens of personen uit een steekproefbestand, waarbij elk record een gewicht heeft dat evenredig is aan de waarschijnlijkheid waarmee het in de populatie thuishoort. Deze gewichten zijn afgeleid van randtotalen of externe marges. Herhaalde trekkingen leveren telkens een nieuwe populatie, die gemiddeld aan de randvoorwaarden voldoet. Deze aanpak is robuust bij beperkte data, maar reproduceert geen correlatiestructuren tenzij expliciet opgelegd. Toepassingen vinden we onder andere in microsimulaties zoals ILUTE (Müller & Axhausen, 2010; Hafezi & Habib, 2014).

- *Bayesiaanse netwerken en Gibbs sampling.* Een meer geavanceerde methode gebruikt conditionele waarschijnlijkheden tussen kenmerken om een populatie op te bouwen. Bayesiaanse netwerken leren de afhankelijkheden tussen variabelen op basis van een steekproef, waarna synthetische entiteiten worden gegenereerd via sampling uit de geschatte joint-distributie. Gibbs sampling is een specifieke Markov Chain Monte Carlo (MCMC)-techniek om dit proces iteratief uit te voeren. Dit is een rekenmethode waarmee steekproeven worden genomen uit complexe kansverdelingen door middel van een opeenvolging van afhankelijke trekkingen. Deze methoden zijn in staat om complexe relaties te modelleren en worden ingezet bij beperkte of onvolledige data (Huynh & Barthélemy, 2018; Rahman & Fatmi, 2023).
- *Combinatorial optimisation sampling (COS).* COS probeert een steekproef samen te stellen uit bestaande records die het best overeenkomt met de gewenste marges, door een optimalisatieproces waarin iteratief records worden verwisseld of herwogen. In tegenstelling tot IPF of entropiemaximalisatie is dit geen mathematisch oplosbaar stelsel, maar een heuristiek. COS combineert de robuustheid van probabilistische sampling met een poging tot structurele consistentie. Borysov et al. (2019) gebruiken deze methode in combinatie met deep generative modelling, waarbij ook synthetische records kunnen worden gegenereerd die niet letterlijk uit een steekproef komen.

Werking
Probabilistische reconstructiemethoden werken volgens kansregels. Waar traditionele methoden streven naar één beste populatie, leveren probabilistische methoden meerdere versies die gemiddeld goed scoren. Sommige methoden, zoals Monte Carlo sampling, zijn relatief eenvoudig en snel, terwijl COS en Bayesiaanse netwerken aanzienlijke rekenkracht en parameterinstellingen vergen. Bij toepassing is het gebruikelijk om meerdere runs uit te voeren en de robuustheid van de resultaten te analyseren.

Toepassingen
Probabilistische reconstructie wordt veel toegepast in agent-based modellen en sociale simulaties (Chapuis et al, 2022). Voorbeelden zijn MATSim (Hörl & Balać, 2020), SimPop, ILUTE en op Bayesiaanse modellen gebaseerde populatie syntheses in Noord-Amerikaanse studies. In onderzoek zijn deze methoden ook benut om populaties te genereren bij incomplete inputdata of bij modellen waarin scenario-onzekerheid wordt gemodelleerd (Chapuis et a;, 2019; Wu & Lyu, 2024).

Voordelen
- Flexibel bij ontbrekende of beperkte input
- Ondersteunt onzekerheidsmodellering en spreiding
- Mogelijkheid om complexe correlatiestructuren te reconstrueren
- Realistisch bij toepassing in agent-based of dynamische modellen

Beperkingen
- Niet-deterministisch: uitkomsten verschillen tussen runs
- Moeilijker te controleren op margeniveau
- Beperkte garantie op plausibiliteit zonder expliciete validatie
- Complexe configuratie bij geavanceerde technieken zoals COS of MCMC

Reflectie op gebruik en validatie
Probabilistische methoden zijn waardevol bij het genereren van alternatieve populaties onder onzekerheid, vooral wanneer deterministische methoden falen of te rigide zijn. Ze bieden flexibiliteit bij beperkte of inconsistente data en zijn goed inzetbaar in agent-based modellen of scenarioverkenningen. Validatie vereist echter een andere aanpak: in plaats van één uitkomst te toetsen, wordt gekeken naar het gemiddelde gedrag en de spreiding over meerdere runs.

Een aandachtspunt is de reproduceerbaarheid. Dit verschilt van deterministische synthesizers, waar dezelfde invoer en instellingen zonder random seeds steeds dezelfde uitkomst geven. Omdat de uitkomsten per run kunnen verschillen, vragen deze methoden om expliciete borging voor toepassingen waarin consistente en herhaalbare resultaten vereist zijn, zoals verkeersprognoses met LMS of NRM. Het gebruik van vaste random seeds kan dit deels ondervangen: bij gelijke invoer levert dezelfde seed dezelfde populatie op. Daarmee wordt technische reproduceerbaarheid bereikt, maar het onderliggende proces blijft stochastisch. Bovendien blijven de uitkomsten gevoelig voor wijzigingen in invoerdata of modelinstellingen. Dat is op zichzelf niet uniek: bij andere methoden verandert de uitkomst ook wanneer invoer of instellingen wijzigen, maar niet tussen identieke runs.

Probabilistische reconstructie vormen vooral een geschikte benadering in verkennende contexten of als onderdeel van hybride systemen. Voor toepassingen waarin beheersbaarheid, stabiliteit en margestructuur centraal staan, zijn deterministische methoden doorgaans beter op hun plaats.

2.3.4 Simulatieve of generatieve modellen

Inleiding
Simulatieve en generatieve modellen bouwen een synthetische populatie niet op basis van randtotalen of herweging van bestaande microdata, maar via een onderliggend model van hoe populaties ontstaan, zich ontwikkelen of zich gedragen. Deze modellen richten zich doorgaans op het proces van populatievorming zelf, en kunnen bijvoorbeeld huishoudens simuleren via gezinsvorming, geboorte, migratie en veroudering, of gedrag genereren op basis van regels of scenario's. Het genereren van populaties is hier dus onderdeel van een dynamisch of gedragsgericht model.

Methoden
- *Rule-based generative simulation.* Deze benadering maakt gebruik van expliciete regels of gedragsmodellen om populaties op te bouwen. Denk aan regels voor samenstelling van huishoudens, leeftijdsopbouw, woningtoewijzing of mobiliteitsgedrag. De simulatie kan tijdstappen bevatten (bijvoorbeeld per jaar), waarbij personen of huishoudens veranderen op basis van probabilistische of deterministische regels. Een voorbeeld is het ALBATROSS-model (Arentze & Timmermans, 2000), dat activiteiten genereert op huishoudniveau en daarvoor een gesynthetiseerde populatie nodig heeft die intern consistent is. In de ALBATROSS-toepassing wordt die populatie doorgaans opgebouwd met iteratieve ophoogmethoden (IPF) en vervolgens als input gebruikt; ALBATROSS is hier dus een voorbeeld van rule-based gedragsgeneratie, niet van een generatieve populatievormingsmethode.
- *Agent-based synthetic population generation.* Hierbij worden individuele 'agents' geprogrammeerd met kenmerken, gedragsregels en interactiemogelijkheden. De

populatie ontstaat als resultaat van de gesimuleerde gedragsinteracties, waarbij het totaalgedrag niet expliciet geprogrammeerd is maar voortkomt uit de samenhang van individuele keuzes. Deze benadering wordt vaak gebruikt in activity-based verkeersmodellen, zoals MATSim (Horni et al., 2016), waarin populaties en activiteiten in één simulatieproces worden opgebouwd. De populatie wordt gegenereerd via gedragsscripts en dus niet direct afgestemd op marges, maar indirect via kalibratie.

- *Dynamic population synthesis.* In sommige toepassingen wordt de populatie niet gegenereerd vanuit een bestaand jaar, maar opgebouwd op basis van scenario's voor toekomstige demografie, huishoudens, economie of mobiliteit. De synthetische populatie volgt dan een exogeen gedefinieerd pad, waarbij bijvoorbeeld veranderingen in arbeidsparticipatie of gezinsgrootte expliciet worden meegegeven (bijv. SimMobility of ILUTE; Miller et al., 2004). Demografische projectie modellen vallen onder dergelijke methoden.

Werking
De populatie ontstaat door simulatie van gedrags- of overgangsregels, al dan niet in tijdstappen. Soms is de populatie een bijproduct van een gedragsmodel; soms is het een expliciete simulatie van populatiedynamiek. Deze aanpak vraagt veel invoer: gedragsregels, overgangskansen, of scenario-parameters. De methoden worden vaak gecombineerd met andere modules, zoals voertuigbezit, activiteitengeneratie of woningmarktmodellen.

Toepassingen
Simulatieve modellen worden vooral gebruikt in omgevingen waarin toekomstige populaties gesimuleerd moeten worden, of waarin de interactie tussen gedrag en context belangrijk is. Toepassingen vinden we in agent-based verkeersmodellen (MATSim, SimMobility), microsimulaties voor ruimtelijke ordening, en langetermijnprognoses waarin maatregelen effect kunnen hebben op populatiekenmerken.

Voordelen
- Flexibel inzetbaar bij toekomstverkenningen
- Leent zich voor integratie met gedragsmodellen en scenarioanalyse
- Maakt het mogelijk om populaties met rijke context en gedrag op te bouwen
- Ondersteunt populatiedynamiek (bijv. gezinsvorming, migratie, veroudering)

Nadelen
- Hoge complexiteit en afhankelijkheid van aannames
- Moeilijker te valideren dan randgedreven methoden
- Beperkte controle over aansluiting op externe marges
- Vereist calibratie, afstemming en doorgaans meer ontwikkeltijd

Reflectie op gebruik en validatie
Simulatieve en generatieve modellen bieden flexibiliteit bij het genereren van populaties waarin gedragskenmerken en context centraal staan. Ze zijn bij uitstek geschikt voor activity-based modellen en langetermijnscenario's, waarbij het gewenst is om veranderingen in huishoudstructuur, arbeidsparticipatie of mobiliteit direct in de populatieopbouw mee te nemen. Voorbeelden zoals MATSim en SimMobility laten zien dat deze benadering goed aansluit bij agent-based simulaties waarin populatie en gedrag samen worden gemodelleerd.

Tegelijkertijd stellen deze methoden hoge eisen aan modelopzet, kalibratie en aannames. De populatie ontstaat niet door directe afstemming op randtotalen, maar als resultaat van een onderliggend model. Dat maakt validatie noodzakelijk, bijvoorbeeld door populatie-uitkomsten te vergelijken met externe verdelingen of door gedragsuitkomsten te toetsen aan enquêtedata. Reproduceerbaarheid is mogelijk bij vaste scenario-invoer, maar bij stochastische processen kunnen uitkomsten per run verschillen.

De ontwikkelkosten, afhankelijkheid van aannames, en complexiteit van validatie maken dat deze benadering minder geschikt is voor toepassingen waarin reproduceerbaarheid, transparantie en aansluiting op bestaande marges belangrijk zijn. Validatie vraagt bovendien extra stappen, omdat de populatie niet direct wordt afgestemd op waargenomen data, maar gegenereerd wordt uit gedragsregels of scenario-invoer.

2.3.5 Datafusie en hybride methoden

Inleiding
Datafusie- en hybride methoden combineren meerdere databronnen en technieken om een synthetische populatie op te bouwen. Deze benadering is ontwikkeld vanuit de constatering dat geen enkele databron volledig is – enquêtes zijn vaak te klein of niet representatief op alle dimensies, terwijl registerdata belangrijke gedrags-kenmerken missen. Door bronnen te combineren (bijvoorbeeld microdata en gedragsdata) en verschillende technieken te integreren (zoals matching, imputatie en machine learning), ontstaat een rijkere en bruikbaardere populatie.

Belangrijkste methode(n)
- *Record linkage en synthetische matching.* Bij record linkage worden datasets samengevoegd op basis van overeenkomende kenmerken (zoals leeftijd, regio of huishoudsamenstelling), al dan niet met probabilistische matching. Dit stelt gebruikers in staat om kenmerken uit verschillende bronnen aan elkaar te koppelen, bijvoorbeeld huishoudkenmerken uit enquête A met mobiliteitsgedrag uit enquête B. Deze techniek is met name relevant als brondata incompleet of partieel zijn (Müller, 2014).
- *Data-imputatie*. Hierbij worden ontbrekende gegevens gesimuleerd of geïmputeerd op basis van statistische of machine learning-methoden, zoals decision trees of regressiemodellen. Deze technieken worden vaak gebruikt in combinatie met IPF of microsimulatie om ontbrekende attributen toe te voegen aan een synthetische populatie. Wu & Lyu (2024) illustreren dit met een methode op basis van diffusion models voor het imputeren van tabulaire data in transport-toepassingen.
- *Deep learning: GANs en VAEs*. Nieuwere benaderingen gebruiken neurale netwerken om realistische synthetische populaties te genereren. Bij Generative Adversarial Networks (GANs) wordt een populatie gegenereerd die niet te onderscheiden is van echte data, door twee netwerken (generator en discrimi-nator) tegen elkaar te trainen . GANs zijn krachtig bij hoge dimensionaliteit en complexe correlatiestructuren (Borysov, Rich & Pereira, 2019). Bij Variational Autoencoders (VAEs) wordt de data eerst gecodeerd naar een lage-dimensionale representatie en vervolgens opnieuw gesampled. Dit maakt het mogelijk om plausibele variaties van de populatie te genereren, ook bij ontbrekende data

(Albiston et al., 2024). Beide technieken worden nog vooral in onderzoeks-omgevingen toegepast, maar bieden perspectief voor toekomstig gebruik – vooral bij privacygevoelige toepassingen, of wanneer klassieke datafusie tekortschiet.

Werking
De populatie wordt opgebouwd door gegevens te koppelen, aan te vullen of geheel synthetisch te genereren. Bij klassieke datafusie gebeurt dit op basis van logica en structuur in de data; bij deep learning-methoden gebeurt dit via een leeralgoritme dat een interne representatie van de populatie leert en daaruit nieuwe eenheden genereert. In hybride toepassingen worden beide gecombineerd, bijvoorbeeld matching gevolgd door GAN-gestuurde aanvulling.

Toepassingen
Datafusie is essentieel wanneer enquêtes niet alle benodigde kenmerken bevatten of wanneer gedrag en achtergrondkenmerken uit verschillende datasets afkomstig zijn. Ook in privacygevoelige contexten, zoals medische of sociale data, is synthetische datageneratie via privacy-preserving technieken in opkomst. GANs en VAEs zijn onder andere in gebruik in Australië en Canada, waar ze gebruikt zijn om privacyveilige populaties te genereren (Albiston et al., 2024; Wu & Lyu, 2024).

Voordelen
- Flexibel inzetbaar bij gescheiden of onvolledige data
- Ondersteunt verrijking van populaties met gedragsdata
- Deep learning-methoden behouden complexe correlatiestructuren
- Privacy-preserving alternatieven voor microdata

Beperkingen
- Matchingkwaliteit bepaalt de bruikbaarheid van de koppeling
- Validatie van synthetische gegevens is complex
- Deep learning vereist veel data, rekenkracht en expertise
- GANs en VAEs zijn nog niet standaard inzetbaar in beleidspraktijk

Gebruik in de praktijk
Datafusie en generatieve technieken bieden krachtige oplossingen voor situaties waarin klassieke methoden tekortschieten – bijvoorbeeld bij datascheiding, onvolledigheid of privacy beperkingen. Binnen transportmodellen kunnen deze technieken nuttig zijn in de voorbereidingsfase of als aanvulling op deterministische methoden. Het werk van Albiston et al. (2024) laat zien dat neurale netwerken synthetische populaties kunnen genereren die intern consistent zijn én vergelijkbare eigenschappen vertonen als waargenomen data. Tegelijkertijd vergen deze technieken validatiebenaderingen die niet alleen naar marges kijken, maar ook naar de structurele plausibiliteit van gegenereerde combinaties. Voor SIVMO zijn dit interessante ontwikkelingen voor de langere termijn, maar voorlopig blijft inzet specialistisch en experimenteel.

2.4 Samenvattende vergelijking van onderzochte methoden

In de voorgaande paragrafen zijn zestien methoden voor populatiesynthese besproken, verdeeld over vijf families. Deze methoden verschillen niet alleen in rekenkundige principes, maar ook in toepasbaarheid, vereisten aan data en gedrag bij implementatie. In tabelvorm vergelijken we de methoden op vijf dimensies die relevant zijn voor verkeers- en vervoersmodellen:

- *Principe*. de rekenkundige of conceptuele grondslag van de methode.
- *Deterministisch.* of de methode bij gelijke input steeds dezelfde uitkomst geeft[1].
- *Complexiteit.* een globale inschatting van rekenkundige en implementatietechnische eisen (laag/middel/hoog).
- *Datavereisten.* het type en detailniveau van inputdata dat nodig is.
- *Toepasbaarheid.* geschiktheid voor verschillende modeltypen: trip-based (Tr), tour-based (To), activity-based (Ac), agent-based (Ag), toekomstscenario's (F)[2].

Nr	Methode	Familie	Deterministisch	Complexiteit	Datavereisten	Toepasbaarheid
1	Iterative Proportional Fitting (IPF)	Iteratieve ophoging	Ja	Laag	Microdata + marges	Tr, To
2	Iterative Proportional Updating (IPU)	Iteratieve ophoging	Ja	Middel	Microdata + marges	Tr, To, beperkt Ac
3	Entropiemaximalisatie	Optimalisatie	Ja	Hoog	Marges (+ prior)	To, Ac, F
4	Least Squares Matching	Optimalisatie	Ja	Hoog	Microdata + marges	Tr, To, Ac
5	KL-divergentie-minimalisatie	Optimalisatie	Ja	Hoog	Microdata + marges	Tr, To, Ac, F
6	Newton-Raphson oplossing (Quad)	Optimalisatie	Ja	Middel–hoog	Microdata + marges	Tr, To
7	Monte Carlo sampling	Probabilistische reconstructie	Nee	Laag	Microdata	Ac, Ag
8	Bayesiaanse netwerken / Gibbs	Probabilistische reconstructie	Nee	Hoog	Microdata	Ac, Ag, verkenningen
9	Combinatorial Optimisation Sampling	Probabilistische reconstructie	Nee	Hoog	Microdata + marges	Ac, Ag
10	Rule-based generative simulation	Simulatief/generatief	Meestal wel	Hoog	Gedragsregels	Ac, Ag, F

[1] Onder 'gelijke input' vallen ook instellingen die de run sturen, zoals een random seed. Een methode is in deze vergelijking 'deterministisch' wanneer zij zonder random component, of bij afwezigheid van een seed, bij identieke invoer en instellingen steeds dezelfde uitkomst oplevert. Methodes die een seed nodig hebben om run-to-run variatie te vermijden, zijn daarmee functioneel stochastisch.

[2] Deze inschatting is gebaseerd op de mate waarin de methode heterogeniteit kan representeren en de populatie-output aansluit op de informatiebehoefte van het betreffende modeltype. 'Toepasbaar' betekent hier: praktisch inzetbaar zonder onevenredige extra aannames of nabewerkng. Een methode die geschikt is voor activity- of agent-based modellen kan in principe ook in trip- of tour-based modellen worden gebruikt, maar kan dan meer detail leveren dan het vraagmodel benut. De tabel duidt daarom vooral de 'passende inzet' aan, niet een harde beperking.

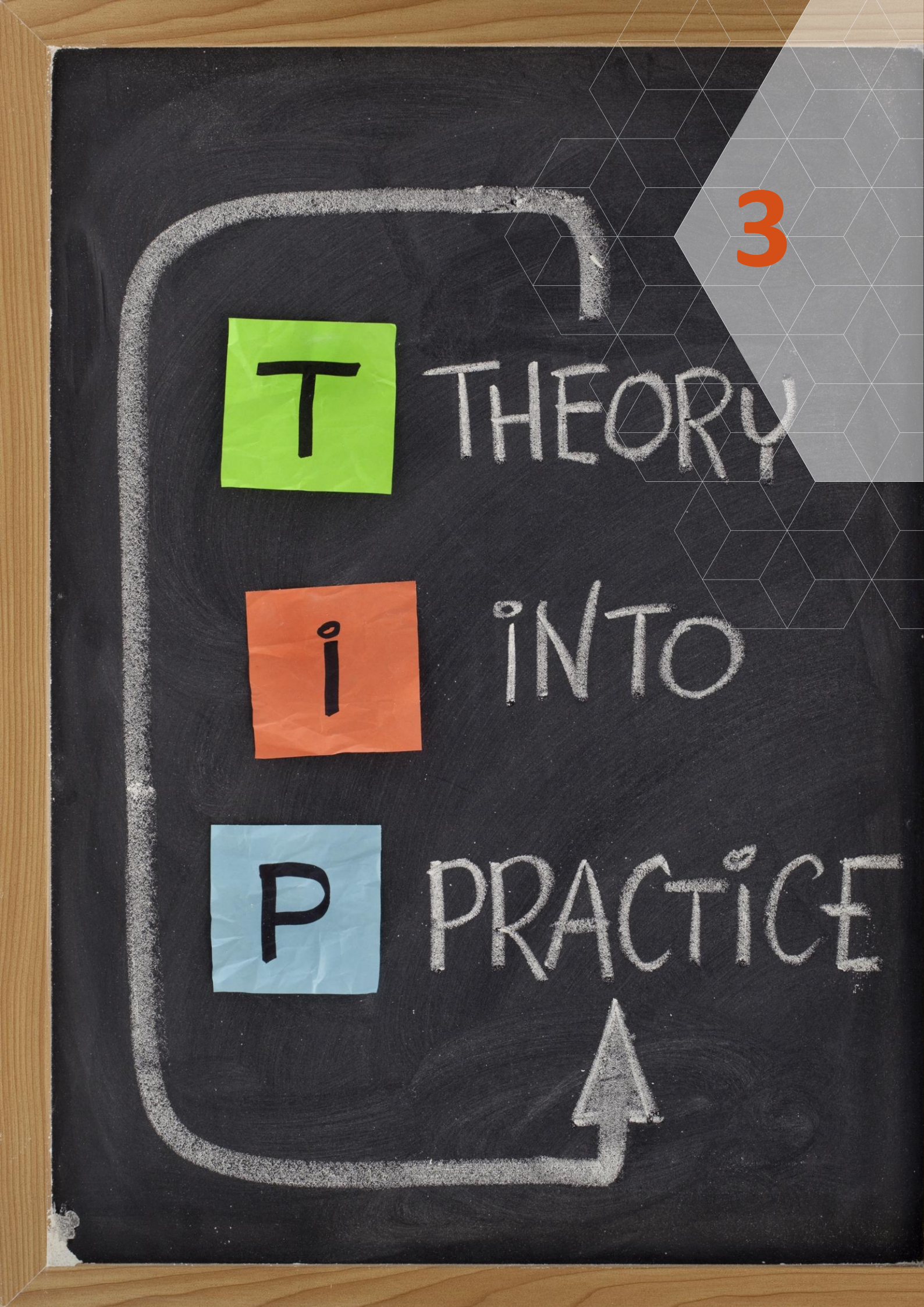
Nr	Methode	Familie	Deterministisch	Complexiteit	Datavereisten	Toepasbaarheid
11	Agent-based populatieopbouw	Simulatief/generatief	Nee	Hoog	Gedragsmodellen	Ac, Ag
12	Dynamic population synthesis	Simulatief/generatief	Ja of nee	Middel–hoog	Scenario-input	F
13	Record linkage	Datafusie / hybride	Ja	Middel	>1 databron	Tr, To, Ac, Ag
14	Imputatie (statistisch/ML)	Datafusie / hybride	Ja of nee	Middel	Microdata	Tr, To, Ac, Ag
15	GANs (Generative Adversarial Nets)	Deep generative models	Nee	Hoog	Microdata	Ac, Ag, verkenningen
16	VAEs (Variational Autoencoders)	Deep generative models	Nee	Hoog	Microdata	Ac, Ag, verkenningen

De vergelijking van methoden laat zien dat populatiesynthese uiteenvalt in een reeks duidelijk te onderscheiden benaderingen, elk met eigen sterktes en beperkingen. Iteratieve methoden zijn transparant en geschikt voor stabiele toepassingen met voldoende microdata en marges, maar missen flexibiliteit bij complexe populatiestructuren. Optimalisatie-gebaseerde methoden bieden meer controle en robuustheid, vooral bij hoge dimensionaliteit of aanvullende randvoorwaarden, maar vereisen nauwkeurige invoer en meer rekenkracht.

Probabilistische reconstructie levert variabele populaties met expliciete onzekerheidsmodellering, en is bruikbaar bij onvolledige data of verkennende simulaties. Voor transportmodellen die reproduceerbaarheid vereisen, is dit echter minder vanzelfsprekend. Simulatieve en generatieve methoden bieden flexibiliteit bij toekomstscenario's en gedragsmodellering, maar vragen om zorgvuldige opzet en validatie. Datafusie- en deep learning-technieken zijn vooral interessant wanneer databronnen incompleet zijn of privacybeperkingen gelden – al bevinden deze zich vaak nog in een onderzoeks- cq ontwikkelfase.

Voor de Nederlandse modelpraktijk geldt dat trip-based en tour-based modellen nog steeds veel worden gebruikt, terwijl over activity-based modellen wordt nagedacht. Binnen SIVMO bestaat het idee om toe te werken naar één generieke populatiesynthese die breed inzetbaar is, dus geschikt voor verschillende modeltypen én schaalniveaus. Dit stelt eisen aan stabiliteit, reproduceerbaarheid, flexibiliteit en schaalbaarheid. Tegen deze achtergrond lijken vooral deterministische en optimalisatie-gebaseerde methoden het meest geschikt als basis. De meer geavanceerde generatieve of probabilistische methoden blijven relevant, maar vragen aanvullende validatie en zijn voorlopig vooral bruikbaar in specifieke contexten. In de volgende hoofdstukken verkennen we welke tools hiervoor in aanmerking komen.

3



3 Tools en software

In dit hoofdstuk bespreken we de belangrijkste software-tools voor populatiesynthese. We belichten hun methodische fundament, toepassingsbereik en praktische relevantie. Na een typologie en vergelijkingstabel gaan we dieper in op geselecteerde tools. We eindigen met een samenvatting uit interviews over gebruiksgemak, onderhoud en toepasbaarheid.

3.1 Typologie en methode

De praktijk van populatiesynthese kent een breed scala aan softwaretools en implementaties. Deze verschillen in methodologische opzet, schaalniveau, openheid en beoogde toepassing. Om structuur te bieden, sluiten we in dit rapport aan bij de indeling in vijf methodologische families zoals uitgewerkt in hoofdstuk 2. Op basis daarvan worden de tools in dit hoofdstuk geordend naar dominante benadering.

We onderscheiden de volgende categorieën van tools:

1. *Tools op basis van iteratieve ophoogmethoden*
   Deze werken met technieken als IPF en IPU, waarbij een steekproefbestand wordt aangepast aan externe marges. Ze zijn transparant, reproduceerbaar en relatief eenvoudig. Voorbeelden:
   – *PopGen* (IPU-gebaseerd, veelvuldig toegepast in de VS)
   – *PopSynWin* (met deterministische en probabilistische modi)
   - *Synthesiser TNO* (IPF gebaseerd)

2. *Tools met een optimalisatie-gebaseerde aanpak*
   In deze groep wordt populatiesynthese benaderd als een optimalisatieprobleem, met technieken als entropiemaximalisatie of minimalisatie van de Kullback-Leibler-divergentie. Deze tools bieden veel flexibiliteit bij complexe randvoorwaarden. Voorbeelden:
   – *SigPopu* (Nederland, KL-divergentie)
   – *Quad* (Nederland, Newton-Raphson methode)
   – *PopulationSim* (VS, entropiemaximalisatie, open source)
   – *SynthPop* (VK, optimalisatie op marges via combinatorisch zoeken)

3. *Tools met probabilistische reconstructie*
   Deze genereren populaties via kansverdelingen of gewichtsgebaseerde sampling. De uitkomsten zijn niet deterministisch, maar gemiddeld consistent met randvoorwaarden. Voorbeelden:
   – *PopSynWin* (in probabilistische modus)
   – *PopGen* (kan ook in samplingmodus worden toegepast)
   – *ILUTE / ILUTS* (Canada, microsimulatie met sampling en gedrag)

4. *Simulatieve of generatieve systemen*
   Hier ontstaat de populatie als resultaat van gedragsmodellen, scenario's of demografische simulaties. Ze worden vaak toegepast in activity-based modellen. Voorbeelden:
   – *SPARK* (Nederland, huishoudvorming en levensloop)
   – *MATSim* (internationaal, agent-based met geïntegreerde populatieopbouw)
   – *ILUTE* (Canada, populatieopbouw als onderdeel van langetermijnsimulatie)

5. *Datafusie-gebaseerde toepassingen met deep learning*
   In deze groep worden synthetische populaties opgebouwd door matching, imputatie of neurale netwerken. De technieken zijn vooral nog in onderzoeksfase. Voorbeelden:
   – *SYNTHETIGAN* (Australië, generative adversarial networks voor synthetische populaties)

Sommige tools combineren methoden (bijv. deterministisch en probabilistisch), of bieden meerdere instelbare algoritmes. In hoofdstuk 3.2 volgt een overzichtstabel met kernkenmerken van de gevonden tools. In hoofdstuk 3.3 worden enkele geselecteerde tools nader beschreven, met aandacht voor structuur, gebruik en toepassingsmogelijkheden in de Nederlandse context.

3.2 Samenvattende vergelijkingstabel

Onderstaande tabel geeft een overzicht van de breedte aan tools die in de literatuur en praktijk zijn aangetroffen. Sommige tools zijn breed inzetbaar en goed gedocumenteerd (zoals *PopulationSim*), andere zijn kleinschalig, in onderzoek of alleen binnen één land of instituut in gebruik (zoals *SigPopu*).

Tool	Methode/familie	Open source	Schaal	Gebruik in praktijk	Opmerkingen
PopGen	IPU / sampling	Nee	Nationaal, regionaal	VS, veelvuldig gebruikt	Beide modi mogelijk (deterministisch, sampling)
PopSynWin	IPF / probabilistisch	Nee	Regionaal	Onderzoek en onderwijs	Eenvoudige GUI, beperkte schaal
SigPopu	Optimalisatie (KL-divergentie)	Nee	Nationaal	Nederland (RWS toepassing)	Vergelijking met Quad
Quad	Newton-Raphson	Nee	Nationaal	Nederland (LMS/NRM)	Stabiliteits-problemen
PopulationSim	Optimalisatie (entropie-maximalisatie)	Ja	Regionaal, nationaal	VS (ActivitySim)	Flexibel en modulair
SynthPop	Optimalisatie	Ja	Regionaal, lokaal	VK (onderzoek)	Gebruikt voor publieke microdata-generatie

Tool	Methode/familie	Open source	Schaal	Gebruik in praktijk	Opmerkingen
ILUTE / ILUTS	Probabilistisch + agent-based gedrag	Nee	Stedelijk	Canada (University of Toronto)	In ontwikkeling, niet direct overdraagbaar
Octavius	Optimalisatie / deterministisch	Nee	Nationaal, egionaal, stedelijk	Almere, Midden-Holland, Drechtsteden	Ontworpen voor stabiele, uitkomsten; integerisatie via SNET i.p.v. sampling
SPARK	Simulatief / levensloop	Nee	Nationaal	Nederland (autobezitmodel)	Demografisch simulatiemodel
MATSim	Simulatief / agent-based	Ja	Nationaal, regionaal	Zwitserland,	Populatie én gedrag in één simulatie
SYNTHETIGAN	Deep learning (GANs)	Gedeeltelijk	Regionaal	Australië (onderzoek)	Privacy-georiënteerd; proof-of-concept

3.3 Toolbeschrijvingen

3.3.1 SigPopu (Nederland)

Achtergrond en methode
SigPopu is ontwikkeld door Significance. De tool minimaliseert de Kullback–Leibler-divergentie tussen een seedverdeling (uit enquêtedata) en externe marges. Dit gebeurt via optimalisatie, waarbij de gesynthetiseerde populatie zowel dicht bij de oorspronkelijke data blijft als voldoet aan randvoorwaarden (Significance, 2024).

Werking
Het proces start met een seed-populatie die gebaseerd is op beschikbare microdata. Vervolgens worden randtotalen (zoals demografische marges) ingevoerd. Een optimalisatie-algoritme zoekt naar gewichten die de KL-divergentie minimaliseren, waarna de populatie wordt gegenereerd met behoud van numerieke consistentie op alle niveaus. Multilevel constraints (bijvoorbeeld per regio en groepskenmerk) worden gecombineerd in één optimalisatieproces.

Toepassing en schaal
SigPopu wordt in Nederland actief gebruikt. Het ondersteunt data op meerdere geografische niveaus, zoals provincies of kleine gebieden. Door gebruik van optimalisatie en integerisatie levert het exacte en reproduceerbare output, zelfs bij uitvoering over meerdere jaren.

Voordelen
- Het heeft een robuuste aanpak, het kan omgaan met lege cellen en minder gangbare combinaties van huishoudens- en persoonskenmerken;
- SigPopu heeft flexibele randvoorwaarden, het is aan te passen met aanvullende marges;

- Het model biedt reproduceerbaarheid, het is deterministisch met consistente resultaten;
- Het kan op meerdere geografische niveaus worden toegepast en in verschillende soorten transportmodellen.

Beperkingen
- Het betreft een gesloten systeem, er is geen open-source code beschikbaar;
- SigPopu vereist expertise voor parameterinstellingen en interpretatie;
- Onderhoud en aanpassing verlopen alleen via de leverancier (Significance).

Validatie in praktijk
In backcasttrajecten voor het LMS werden gesynthetiseerde verdelingen (bijvoorbeeld huishoudsamenstelling en auto- of OV-bezit) gevalideerd door vergelijking met geobserveerde data over meerdere jaren. De tool toonde goede marge-fit en stabiele output, met uitzondering van enkele kleine subgroepen.

Beschikbaarheid en toegankelijkheid
SigPopu is een tool van Significance. De benodigde expertise omvat kennis van statistische optimalisatie en kennis van Nederlandse datasets.

Reflectie op gebruik binnen SIVMO
SigPopu is geschikt voor SIVMO-toepassing, vooral als onderdeel van trip-based en tour-based modellen op regionaal of nationaal niveau. De combinatie van robuuste methode en reproduceerbaarheid is aantrekkelijk. Tegelijkertijd vormt het gesloten karakter een risico voor onderhoud, kennisdeling en inpassing in een open infrastructuur. Voor SIVMO zou een meer open alternatief of een hybride model de voorkeur kunnen hebben.

3.3.2 Quad (Nederland)

Achtergrond en methode
Quad is een module die is geïntegreerd in LMS, NRM en Venom. Quad past de Newton–Raphson-methode toe voor gewichtsoptimalisatie, zodat de synthetische populatie voldoet aan opgelegde marges, op een snelle en wiskundige manier via de oplossing van een stelsel niet-lineaire vergelijkingen .

Werking
De tool gebruikt het Newton–Raphson-algoritme om gewichten toe te passen op microdata (ODiN), zodat ze nauw aansluiten op randtotalen. De gewichten worden zo geoptimaliseerd dat de populatie marges exact matcht. Door de gewichten iteratief te aan te passen, convergeert de oplossing snel, mits de aannames en initiële waarden adequaat zijn.

Toepassing en schaal
Quad wordt al jarenlang toegepast binnen het LMS- en NRM voor uiteenlopende vervoers- en verkeersstudies. Quad functioneert effectief bij brede databundels, maar ‘loopt vast’ bij kleine huishoudenstypen die weinig of geen representatie in de seed-data hebben.

Voordelen

- Quad is snel, een convergentie wordt bereikt binnen enkele iteraties en het is is effectief voor grote datasets;
- Het is deterministisch en reproduceerbaar, identieke invoer geeft identieke uitkomsten.

Beperkingen

- Quad is instabiel bij zeldzame combinaties. Kleine of lege categorieën kunnen leiden tot sterke fluctuaties of divergent gedrag in gewichten;
- Het is gevoelig voor de startsituatie, de Newton–Raphson-methode vereist goede initiële waarden en goede randvoorwaarden om te convergeren

Validatie in praktijk

Tijdens testvergelijkingen met SigPopu bleek Quad snelle resultaten te leveren, maar vaak met onrealistische uitkomsten voor zeer kleine huishoudenstypen. Bij dergelijke categorieën traden instabiliteit en extreme ophoogfactoren op, wat vragen opriep over de betrouwbaarheid van kleine segmenten in landelijke toepassingen.

Beschikbaarheid en toegankelijkheid

Quad is ingebouwd in het LMS en NRM en wordt beheerd vanuit Rijkswaterstaat. De broncode is eigendom van Rijkswaterstaat, maar niet publiek beschikbaar; toegang en support verlopen via Rijkswaterstaat.

Reflectie op gebruik binnen SIVMO

Quad is geschikt als snelle tool voor trip- en tour-based modellen waar robuuste oplossingen op grote schaal gewenst zijn. Door het inherente risico van instabiliteit bij minder gangbare segmenten en de beperkte toegankelijkheid is het echter niet geschikt als generieke synthesetool voor alle modeltypen. Voor toepassing in modellen van de SIVMO partners is een alternatief zoals SigPopu of een open-source methode robuuster en meer toekomstbestendig.

3.3.3 Octavius (Nederland)

Achtergrond en methode

Octavius is ontwikkeld door Goudappel en DAT.mobility en wordt toegepast in regionale modellen in Nederland. De population synthesizer gebruikt een deterministische aanpak zonder sampling, gericht op stabiele en vergelijkbare uitkomsten bij gelijke invoer. De methode sluit aan bij een tweedeling tussen fitting (aansluiten op marges) en allocation (consistente samenstelling van personen en huishoudens). De keten bestaat in de praktijk uit IPF voor afzonderlijke persoons- en huishoudverdelingen, een (iterative) non-negative least squares stap (NNLS/INNLS) voor samenstelling, en SNET voor deterministische integerisatie.

Werking

Octavius start met een seed-steekproef (zoals ODiN en MPN) en marginale gegevens per zone, vaak gebaseerd op CBS-buurtstatistieken en projectafhankelijke aanvullingen. Per zone worden met IPF de verdelingen voor persoonssegmenten en huishoudsegmenten gefit op de opgelegde marges. Daarna combineert INNLS de persoons- en huishoudverdelingen tot concrete huishoudsamenstellingen, zodat persoonskenmerken en huishoudkenmerken onderling consistent zijn. De continue

oplossing wordt vervolgens met SNET omgezet naar discrete personen en huishoudens, zonder toevalsprocessen.

Omdat CBS-microdata buiten de RA-omgeving niet gebruikt kunnen worden, worden afhankelijkheden tussen kenmerken in de praktijk gereconstrueerd uit gecombineerde bronnen. Soms gebeurt dit via imputatie met oudere, grotere steekproeven zoals OVG of MON.

Toepassing en schaal
Octavius is onder meer ingezet of in ontwikkeling en toepassing voor Almere, Zwolle, Midden-Holland, Drechtsteden en Purmerend. De synthesizer kan op meerdere schaalniveaus draaien, maar de haalbare resolutie hangt af van de beschikbaarheid en consistentie van marges. Synthese op een niveau fijner dan PC4 vraagt extra marges op dat niveau.

Voordelen
- Reproduceerbare uitkomsten bij gelijke invoer, doordat sampling wordt vermeden.
- INNLS koppelt persoons- en huishoudverdelingen expliciet, wat consistente huishoudsamenstellingen ondersteunt.
- SNET zet continue uitkomsten deterministisch om naar discrete aantallen, met beperkte en stabiele discretisatiefout.
- Geschikt voor scenariovergelijkingen doordat ruis tussen runs wordt beperkt.
- Modulair opgezet en koppelbaar via interfaces aan verschillende omgevingen.

Beperkingen
- Resultaten zijn gevoelig voor de volledigheid en consistentie van marges, vooral bij fijnmazige zones en rijke kruistabellen.
- Bij zeldzame combinaties of lege cellen zijn pragmatische ingrepen nodig (samenvoegen/uitsluiten), wat margeschendingen in subgroepen kan geven.
- Afhankelijkheden tussen kenmerken moeten vaak indirect worden gereconstrueerd uit gecombineerde bronnen; dit kan beperkingen geven bij correlaties.
- Externe validatie met CBS-microdata wordt niet als standaardpraktijk genoemd, waardoor validatie vooral op margecontrole leunt.
- Niet openbaar en licentiegebonden; gebruik, onderhoud en doorontwikkeling hangen af van de leverancier.

Validatie in praktijk
Octavius is niet extern gevalideerd met CBS-microdata. Interne validaties komen wel voor, bijvoorbeeld bij kalibratie in regionale projecten. In de praktijk domineert margecontrole, terwijl toetsing aan externe observaties beperkt blijft tot verkennende analyses of expertbeoordeling.

Beschikbaarheid en toegankelijkheid
Octavius is ontwikkeld bij Goudappel en DAT.mobility. De synthesemodule is niet openbaar, het gebruik en onderhoud zijn licentiegebonden. In het interview wordt de implementatie als een zelfstandige Java-module beschreven, die via interfaces aan platformen kunnen koppelen.

Reflectie voor SIVMO
Octavius past bij toepassingen waar scenariovergelijking en reproduceerbaarheid zwaar wegen, doordat de population synthesizer deterministisch is en integerisatie zonder toeval plaatsvindt. Voor SIVMO is de beperkte openbaarheid een aandachtspunt, zowel voor transparantie als voor bredere overdraagbaarheid en doorontwikkeling.

Als Octavius wordt overwogen voor bredere inzet, ligt een samenwerkingsroute met de leverancier het meest voor de hand, aangevuld met expliciete afspraken over documentatie, validatie en beheer

3.3.4 PopulationSim (VS)

Achtergrond en methode
PopulationSim is ontwikkeld door Transport Foundry en de Oregon Department of Transportation, als onderdeel van het open-source ActivitySim-platform voor activity-based modellen (GitHub). De tool baseert zich op entropiemaximalisatie, ook wel list balancing genoemd, om synthetische populaties te genereren die voldoen aan randvoorwaarden terwijl ze minimale extra structuur aannemen (Paul et al, 2018). Voor integerisatie wordt gebruikgemaakt van lineaire programmering, zodat gewichten worden omgezet in volledige huishoud- en personenrecords.

Werking
PopulationSim vereist twee inputbronnen: een seed-steekproef (bv. PUMS, in Nederland ODiN) en marginale data op meerdere geografie-niveaus (TAZ, counties od postcodes in Nederland). Zie activitysim.github.io voor een gedetailleerde beschrijving. Daarna volgt:

1. Entropiemaximalisatie met list balancing om gewichten te optimaliseren.
2. Simultane integerisatie via lineaire programmering om volledige entiteiten te creëren.
3. De tool vermijdt “sampling zeros” en propagatiefouten (meetfouten doorgeven aan het eindresultaat) vanwege gelijktijdige balans over regio’s .

Toepassing en schaal
PopulationSim wordt actief gebruikt door Oregon DOT en MWCOG voor grootschalige synthese van huishoud- en personenpopulaties op verschillende geografische niveaus. De tool is getest en gevalideerd op gebieden met miljoenen inwoners en hoge dimensionaliteit (zie bijvoorbeeld RSG, 2021).

Voordelen

- Het model is open source en transparant. Het is beschikbaar via GitHub met BSD-licentie.
- PopulationSim geeft reproduceerbare resultaten en is stabiel. Het gebruikt deterministische methoden met integerisatie en geeft consistente uitkomsten.
- Het is robuust bij sampling-zeros en fouten in kleine zones door simultane balans (Paul et al, 2018).
- Het model is flexibel qua schaal en steekproefdimensies en daarmee geschikt voor uiteenlopende geografische en attributieniveaus .

Beperkingen

- Het model is afhankelijk van goede margedata. Ontbrekende of inconsistente controles verslechteren resultaten.
- Integerisatie heeft kalibratie nodig om grote afwijkingen in gewichten te voorkomen.
- Er is sprake van variabiliteit bij beperkte runtime. Bij time-outs in integerisatie kunnen de disaggregate resultaten licht variëren.

Validatie in praktijk

PopulationSim biedt uitgebreide validatiefuncties, inclusief rapportages over gemiddelde afwijkingen, SD en RMSE op verschillende geografische niveaus. Validaties bij MWCOG tonen nauwkeurige marges en evenmatige verdeling van expansion-factors (RSG, 2021).

Beschikbaarheid en toegankelijkheid

De tool is publiek beschikbaar via de ActivitySim GitHub-repository met code, documentatie en voorbeelden (zie GitHub). Er is een Python-installatie beschikbaar, inclusief scripts van RSG voor setup en validatie (zie bijvoorbeeld RSG 2021).

Reflectie op gebruik binnen SIVMO

PopulationSim biedt een goede basis voor één generieke synthesetool voor trip-, tour- en activity-based modellen. De combinatie van openheid, reproduceerbaarheid, schaalbaarheid en robuustheid sluit goed aan bij de eisen die Nederlandse transportmodellen stellen. Aandacht is nodig voor:

- Aansluiting op CBS-datasets en de marginale structuur.
- Zorgvuldige keuze van integerisatieparameters.
- Integratie van de tool in Nederlandse modellen.

PopulationSim kan dienen als goede populatie synthesizer voor transportmodellen in Nederland.

3.3.5 PopGen (VS)

Achtergrond en methode

PopGen is een populair hulpmiddel voor populatiesynthese, ontwikkeld door de Mobility Analytics Research Group (MARG) aan de Arizona State University, met ondersteuning van de US Federal Highway Administration Link). De kern van de tool is het Enhanced Iterative Proportional Updating (IPU)-algoritme, een geavanceerde variant van IPF die zowel huishoud- als persoonsattributen tegelijkertijd afstemt op marginale totalen. Deze methode lost daarmee een bekend knelpunt op van traditionele IPF-systemen, waar persoonskenmerken vaak niet adequaat geregeld worden.

Werking

PopGen past multi-level IPU toe, waarmee voor verschillende geografische lagen tegelijk een representatieve populatie kan worden gegenereerd. Het proces verdeelt gewichten vanuit een seed-surveybestand over zones, gevolgd door herhaalde herwegingen totdat zowel huishouden- als persoonsspecifieke marges kloppen. De tool ondersteunt bovendien heuristische optimalisaties en probabilistische simulatie

om zero-cell problemen aan te pakken, veroorzaakt door combinaties van kenmerken die niet in de sample voorkomen.

Toepassingen en schaal
PopGen is beproefd in regio's zoals Maricopa County (Arizona), Greater Toronto (als onderdeel van ILUTE), en gebruikt binnen Metropolitan Planning Organizations in de VS. Het proces toont zich efficiënt, ook bij populaties van meerdere miljoenen agents zoals in New York City, waar generatie van ~8 miljoen populanten binnen 12 minuten lukte op standaard hardware .

Voordelen
- Robuust beheer van multi-level marges.
- Handelbaar in grote datasets.
- Open source toegang tot IPU-algoritme via MARG

Beperkingen
- Vereist nauwkeurige margedata op meerdere niveaus.
- Heuristische optimalisatie voegt complexe parameterafstemming toe.
- Mogelijk closed/partly-open code, afhankelijk van versie.

Validatie in praktijk
In studies zoals Kagho et al. (2020) toont PopGen near-perfecte marge-passingen (R^2 ≈ 0,999) voor zowel totaalpopulatie als persoons/huishoudvariabelen binnen kleine zones — met afwijkingen slechts enkele procenten in leeftijdscategorieën. Dergelijke resultaten bevestigen de betrouwbaarheid van PopGen voor microsimulatiemodellen.

Beschikbaarheid en toegankelijkheid
PopGen 2.0 is beschikbaar via de MARG-website. De tool wordt ondersteund door MARG, met documentatie, methodologische toelichting en toepassingsvoorbeelden.

Reflectie
PopGen biedt een goede basis voor trip- en tour-based modellen: de precisie, reproduceerbaarheid en schaalbaarheid zijn sterke troeven. Voor activity-based toepassingen is de stap naar microsimulatoren aanwezig. De open-source aard ondersteunt onderhoud en aanpassing, maar een zorgvuldige kalibratie is belangrijk. Binnen de SIVMO context is het van belang dat de toetsing op Nederlandse marges, performance in regionale schaal en integratie met bestaande modellen verder worden onderzocht..

3.3.6 PopSynWin (VS)

Achtergrond en methode
PopSynWin is ontwikkeld in 2008 door de Universiteit van Illinois in Chicago (UIC) als een stand-alone synthesetool voor populaties op huishoud- en persoonsniveau. De kern van de methode is IPF, waarbij steekproefdata (seed data) herhaaldelijk worden aangepast om te voldoen aan bekende marginale totalen bijvoorbeeld verkregen uit censusdata. Het uitgangspunt is dat de resulterende microdata de oorspronkelijke correlatiestructuur behouden en minimale extra informatie introduceren.

Werking
De methode opereert via iteratieve aanpassingen van celwaarden in kruistabellen, totdat zowel rijen- als kolomtotalen overeenkomen met opgelegde marges. Naast standaard IPF introduceert PopSynWin automatische optimalisatie van categorisatie-detaillering en past het gewichten toe om persoonsniveau-voorwaarden tegelijk te garanderen (Müller et al, 2010). Voor grotere attributensets hanteert de tool geheugenoptimalisaties via sparse datarepresentatie.

Toepassing en schaal
PopSynWin is gevalideerd voor synthese van de Chicago-populatie en werd in 2015 gebruikt voor Greater Melbourne. De tool biedt zowel IPF- als probabilistische modi via een grafische interface. In vergelijking met PopGen bleek IPF nuttig voor het nauwkeurig opbouwen van huishoudniveau-marges, terwijl IPU beter presteerde bij persoonskenmerken (Jain et al, 2015).

Voordelen
- Interface met zowel IPF als probabilistische opties
- Behoudt de correlatiestructuur uit seed data
- Automatische aanpassing van categorisatiedetaillering
- Herbruikbaar met data uit verschillende geografische schaalniveaus

Beperkingen
- Moeilijk schaalbaar bij hoge dimensionaliteit wegens geheugen- en convergentielimieten
- Zero-cell issues blijven problematisch bij minder gangbare combinaties van huishouden- en persoonskenmerken.
- Beperkte gebruikers community en minder geschikt voor grootschalige nationale toepassingen

Validatie in praktijk
Case studies tonen goede marge-fit op huishoudeniveau (bijv. Melbourne), maar benadrukken dat de tool minder geschikt is voor persoonsniveau zonder aanvullende technieken (Müller et al, 2010). Validatiemethoden omvatten aggregatie van microdata en consistente margevergelijking tegen censusgegevens.

Beschikbaarheid en toegankelijkheid
PopSynWin is open-source beschikbaar via deze Link . Gebruikerservaringen laten zien dat de GUI-dialoog gemakkelijk inzetbaar is voor pilotprojecten, maar er is slechts beperkte online documentatie en actieve ondersteuning.

Reflectie op gebruik binnen SIVMO
PopSynWin is geschikt als piloottool of educatief instrument vanwege eenvoud en visualisatie via GUI. Voor kleinschalige of onderwijsdoeleinden biedt het een toegankelijke kennismaking met IPF-methodologie. Voor algemene toepassing in Nederlandse modellen zijn schaalbaarheid, performance en ondersteuning echter beperkt. Voor grootschalige of intensieve toepassing heeft een tool met bredere schaalcapaciteit, API-ondersteuning en actieve community, zoals PopulationSim of PopGen, de voorkeur.

3.3.7 MATSim population synthesizer (Switzerland)

Achtergrond en methode
MATSim (Multi-Agent Transport Simulation) is een open-source agent-based mobiliteitsmodel geschreven in Java. Het simuleert beslissingsgedrag en interacties van individuele agenten (personen/huishoudens) over de dag heen, om zo mobiliteitspatronen te voorspellen en te analyseren (https://matsim.org/). Populatiesynthese is hierbij essentieel, aangezien volledige microdata vaak niet beschikbaar zijn.

MATSim bevat een module voor populatiesynthese, de zogenaamde PoPulationGenerator. Deze genereert een basispopulatie door records (personen) te creëren op basis van externe data (zoals volkstelling of enquêtes). De module maakt gebruik van data-gedreven algoritmen. Er is momenteel geen geïntegreerde IPF/IPU-module binnen MATSim; implementaties gebruiken vaak kleine Java-scripts of hulpmiddelen zoals node-matsim-population-generator als proof-of-concept.

Werking
Bij uitvoering van de Population Generator binnen MATSim start het systeem met beperkte input, zoals demografische data uit volkstellingen of enquêtes. Op basis daarvan wordt een plausibele populatie opgebouwd: elk individu krijgt demografische kenmerken (zoals leeftijd en woonlocatie) en wordt gekoppeld aan een gestructureerd reisschema – thuis, werk, winkel, terug naar huis.

De methode zorgt voor reproduceerbaarheid: door gebruik van een vaste ‘seed’ levert elke run dezelfde populatie op. De gegenereerde populatie wordt opgeslagen in gangbare formaten (o.a. CSV/XML), die later kunnen worden ingelezen door MATSim voor activiteitensimulaties.

Toepassing en schaal
MATSim wordt wereldwijd ingezet, onder meer voor stedelijke en regionale simulaties (bijv. Nouakchott, Chicago, Melbourne). Het is geschikt voor populaties van duizenden tot enkele miljoenen agents, afhankelijk van beschikbare gegevens en scripts.

Voordelen
- MATSim is volledig open-source, de code is beschikbaar via GitHub, inclusief voorbeelden.
- Het systeem is flexibel en uitbreidbaar, de gebruikers kunnen eigen logica toevoegen in Java of via externe scripts.
- Het resultaat is reproduceerbaar, onder meer door gebruik van vaste random seeds.
- De populatie synthese is geïntegreerd met simulatiemodel, de populatie vormt directe input voor gedragsplanning.

Beperkingen
- MATSim heef geen statistisch geavanceerde synthese: er is bijvoorbeeld geen IPF/IPU. De validatie en re-weighting worden extern opgezet.
- Er is geen ingebouwde marge-controle; dit moet via corrigerende scripts extern worden geregeld.

- Het model is technisch complex. Bij ontwikkeling vereist het kennis van Java, het is minder geschikt voor modellen met beperkte IT-ondersteuning.

Validatie in de praktijk
MATSim biedt geen ingebouwde validatietools voor populatie versus marges. Validatie wordt doorgaans uitgevoerd via externe analyses: populatie wordt geaggregeerd (bijv. naar leeftijdsgroepen of zones) en vergeleken met census- of surveydata. Geïntegreerde routines ontbreken, waardoor reproduceerbaarheid afhankelijk is van scripts en documentatie.

Beschikbaarheid en toegankelijkheid
De Java-implementatie is openbaar beschikbaar via MATSim GitHub en documentatie. Er bestaan voorbeelden (RunZPopulationGenerator) die basispopulatiegeneratie demonstreren (zie MATSim).

Reflectie voor SIVMO
MATSim's populatiesynthese biedt sterke integratie met agent-based modellering en reproduceerbaarheid via code en seeds. Voor SIVMO partners is het in de toekomst geschikt als onderdeel van een activity-based modellen. Dit vereist wel verder onderzoek. Vooralsnog is het voor eenvoudigere trip- of tour-based modellen niet geschikt.

3.3.8 SynthPop (VK)

Achtergrond en methode
SynthPop is een open-source R-pakket ontwikkeld door Nowok, Raab en Dibben (Nowok et al, 2016) aan de University of Edinburgh. De tool genereert synthetische microdata op basis van publieke datasets door variabelen één voor één te modelleren met regressie- of boommodellen (bijvoorbeeld CART). Vervolgens wordt sequentiële conditionele synthese toegepast om onderlinge relaties te behouden en tegelijk privacybescherming te waarborgen (synthpop.org.uk).

Werking
SynthPop vervangt waarden in een dataset door nieuwe, synthetisch gegenereerde waarden, met behoud van de statistische kenmerken van de oorspronkelijke gegevens. De gebruiker kan kiezen uit verschillende methoden per variabele en de sequentie waarin variabelen gesynthetiseerd worden, configureren via de functie syn() (Nowok et al, 2016).

Toepassing en schaal
De tool wordt vooral gebruikt binnen het VK door statistiekbureaus zoals de Scottish Longitudinal Study (SLS), maar kent ook internationaal gebruik bij academische instellingen . De schaal varieert van datasets met duizenden tot tienduizenden records, zelfs honderden variabelen.
Voordelen
- Open-source en configureerbaar via R.
- Ondersteuning voor verschillende synthesemethoden (CART, logistische regressie, random forests).
- Ingebouwde tools voor utility- en disclosure-analyse, inclusief.

Beperkingen

- Validatie en privacycontrole zijn voornamelijk academisch en minder toegepast in operationele vervoersdomeinen.
- Gebruikers moeten parameters zorgvuldig instellen: defaults zijn niet altijd voldoende voor specifieke datasets.

Validatie in praktijk

SynthPop bevat routines voor validatie zoals propensity scores en afstandsmetingen. De ontwikkelaars tonen aan dat synthetisch en origineel modelresultaten statistisch overeenkomen, bijvoorbeeld in cohortanalyses.

Beschikbaarheid en toegankelijkheid

- Website: synthpop.org.uk
- Beschikbaar via CRAN: https://cran.r-project.org/package=synthpop

Reflectie op gebruik binnen SIVMO

SynthPop is een interessant open-source alternatief wanneer gewerkt wordt met publieke data en privacy belangrijk is. Vooral geschikt voor het voorbereiden van datasets met uitgebreide validatie. Voor SIVMO kan SynthPop een waardevolle aanvulling zijn tijdens de verkennende onderzoeksfase. Voor operationeel gebruik binnen vervoersmodellen zijn aanvullende aanpassingen nodig, zoals koppeling met marges, reisgedragskenmerken en integratie met modelstructuren. Vooralsnog als generieke tool minder geschikt.

3.3.9 Overige synthesizers

SYNTHETIGAN (Australië; prototype)

SYNTHETIGAN is een experimenteel populatiesyntheseproject dat gebruikmaakt van Generative Adversarial Networks (GANs). Deze methode bestaat uit een generator en een discriminator die tegen elkaar 'spelen'. De generator leert om synthetische populaties te creëren die voor de discriminator niet van echte data te onderscheiden zijn, waardoor complexe correlaties tussen attributen zoals leeftijd, huishoudtype en voertuigbezit behouden blijven (Albiston et al., 2024).

Het project bevindt zich nog in een onderzoeksfase—gericht op privacy vriendelijke populaties met rijke structurele kenmerken. Voordelen zijn de potentie om geavanceerde correlatiestructuren te modelleren en persoonlijke gegevens te beschermen. Nadelen zijn de hoge rekenlast, experimentele status en beperkte beschikbaarheid als werkend prototype. SYNTHETIGAN is nog niet gebruiksklaar voor structurele implementatie in modellen zoals bij SIVMO partners.

ILUTE / ILUTS (Canada)

ILUTE (Integrated Land-Use, Transportation and Environment) is ontwikkeld aan de University of Toronto. Het combineert populatie- en huishoudopbouw met landgebruik en transportmodellen in één microsimulatie. Voor de populatiesynthese gebruikt het een lijst-gebaseerde variant van IPF, waarbij individuele records systematisch worden aangepast aan marges (Miller & Salvini, 2005) .

ILUTE is toegepast in de Greater Toronto Area voor langetermijnscenario's met veranderende demografie en huishoudstructuur. De aanpak is efficiënt in geheugengebruik en handhaaft talrijke attributen. Tegelijk is het model complex en sterk geïntegreerd in ILUTE, wat het lastig maakt om als losstaande tool te gebruiken. De methode illustreert wel hoe uitgebreide integratie van technieken mogelijk is in grootschalige toepassingen.

SPARK (Nederland)
SPARK is opgezet als dynamisch autobezitsmodel waarin leefpatronen en huishoudkenmerken in de loop der tijd veranderen. Hoewel het primair gericht is op autobezit, bevat het methoden die ook voor populatiesynthese toegepast zouden kunnen worden.

SPARK is nooit structureel ingezet voor populatiesynthese. Het zou geschikt zijn in stimulaties van levensloopveranderingen, maar als zelfstandig synthese-instrument is het minder ontwikkeld. Voor SIVMO biedt SPARK op dit moment beperkte meerwaarde buiten de autobezit context.

PopSynth (Canada)
PopSynth is een tool die populaties genereert via heuristische bewerkingen—zoals het wisselen, toevoegen of verwijderen van huishoudrecords—geleid door een 'goodness-of-fit'-score. Deze aanpak gebruikt zowel deterministische als heuristische optimalisatiemechanismen om margeovereenstemming te bereiken.

De tool is toegepast in enkele Noord-Amerikaanse stedelijke modellen en blijkt flexibel bij multilevelrandvoorwaarden. Echter, het is commercieel en heeft een beperkte community. Voor gebruik is een licentie vereist en aanpassing aan lokale data. PopSynth illustreert wel een hybride aanpak waarbij heuristiek met optimalisatie wordt gecombineerd in populatiesynthese.

Click here and insert picture via Insert Picture. To enlarge: click in corner and drag. To move: click on Crop on right Format tab.

4

4 Bevindingen uit interviews

Dit hoofdstuk beschrijft de bevindingen uit de interviews met organisaties die populatiesynthese ontwikkelen, toepassen of randvoorwaarden stellen aan datagebruik. De interviews vullen het literatuuronderzoek aan door inzicht te geven in praktische keuzes, beperkingen en werkwijzen, met nadruk op datatoegang, privacy, schaalniveau, validatie en beheer. De uitkomsten laten zowel overeenkomsten als verschillen zien tussen ontwikkelaars, gebruikers en databeheerders, en vormen een basis voor de latere beoordeling van methoden en tools in dit rapport.

4.1 Overzicht van geïnterviewde partijen

Voor deze studie zijn zes interviews afgenomen met organisaties die ieder op een eigen wijze betrokken zijn bij populatiesynthese. De selectie is gericht op partijen die ervaring hebben met het ontwikkelen van synthesizers, het toepassen ervan in vervoers- en verkeersmodellen, of het beheren van de onderliggende databronnen. Daarmee vertegenwoordigen de gesprekken een breed spectrum van methodologische, organisatorische en datagerelateerde perspectieven. Dit overzicht biedt context voor de meer gedetailleerde beschrijvingen in de volgende paragrafen.

PTV is een internationale ontwikkelaar van modelsoftware, waaronder Visum. In Visum is een geïntegreerde populatiesynthesemodule opgenomen die gebruikt wordt in diverse landen. PTV geeft inzicht in methodologische keuzes, onderhoud en schaalbaarheid van synthesizers binnen een commercieel softwareplatform (zie ook bijlage 3.1)

Department for Transport (DfT) in het Verenigd Koninkrijk ontwikkelt een nationale populatiesynthese die wordt ingezet voor transportmodellen op verschillende schaalniveaus. DfT geeft daarmee een kader voor de organisatie van syntheses, kwaliteitsborging en governance in een nationale context (zie ook bijlage 3.2).

TNO ontwikkelt populatiesyntheses voor gebruik in een activity-based model. De aanpak van TNO is gebaseerd op iteratieve ophoogmethoden en maakt gebruik van CBS-microdata. Daarmee biedt TNO inzicht in de technische en data-gerelateerde uitvoerbaarheid binnen de Nederlandse context (zie ook bijlage 3.3).

Centraal Bureau voor de Statistiek (CBS) is beheerder van de microdatabronnen die de basis vormen voor populatiesyntheses in Nederland. Het CBS heeft specifieke voorzieningen, privacykaders en voorwaarden voor het gebruik van deze data. Deze randvoorwaarden bepalen mede de schaal en toepasbaarheid van synthesizers (zie ook bijlage 3.4).

Goudappel is ontwikkelaar van de Octavius, een populatiesynthesizer binnen Omnitrans. Deze synthesizer wordt ingezet in regionale modellen in Nederland. Het interview geeft inzicht in praktische toepassing, validatie en onderhoud binnen een operationele modelketen (zie ook bijlage 3.5).

Significance ontwikkelt en onderhoudt SigPopu en is beheerder van Quad in het LMS. Deze synthesizers wordt toegepast in nationale studies en modellen en bieden inzicht in methodologische stabiliteit, databehoefte en toekomstige doorontwikkeling (zie ook bijlage 3.6).

De interviews met deze zes partijen bieden gezamenlijk een beeld van de huidige praktijk en de relevante technische en organisatorische aandachtspunten. In de volgende paragrafen worden de bevindingen per organisatie beschreven.

4.2 PTV

Achtergrond en rol
PTV ontwikkelt en onderhoudt het modelplatform Visum, waarin vanaf versie 2026 een geïntegreerde populatiesynthesemodule beschikbaar is. Deze vervangt de eerdere afhankelijkheid van de externe tool PopulationSim, die door veel gebruikers moeilijk te installeren en te onderhouden was. Problemen met Python-dependencies, beperkte rechten binnen overheidsomgevingen en instabiele updates maakten dat PopulationSim in de praktijk vaak niet werkte. PTV beschouwt de integratie in Visum daarom als een pragmatische, noodzakelijke stap: niet gedreven door methodologische vernieuwing, maar door de behoefte aan een stabiel, reproduceerbaar en breed toepasbaar systeem dat “voor iedereen werkt, niet alleen voor specialisten”.

De geïntegreerde synthesizer vormt nu een vaste component binnen Visum en sluit aan op diverse modeltypen, waaronder trip-based, tour-based en agent-based modellen. Binnen PTV’s eigen workflow (o.a. Model2Go) is populatiesynthese de eerste stap in een automatische modelketen, waarmee PTV wil bereiken dat synthese een routinematige, foutloze en herhaalbare bewerking wordt.

Methode
De populatiesynthese in Visum is methodologisch nauw verwant aan PopulationSim. De module maakt gebruik van een entropiebenadering (“list balancing”), waarbij continue gewichten worden geschat zodat huishoudens en personen gezamenlijk aansluiten op opgelegde marges. De integerisatie gebeurt via een lineaire programmastap, die ervoor zorgt dat er discrete huishoudens en personen ontstaan zonder ophoping van afrondingsfouten[3].

[3] Concreet betekent dit dat de continue gewichten worden omgezet naar hele aantallen door een optimalisatieprobleem op te lossen, waarbij per huishouden of persoon wordt bepaald of deze 0, 1, 2, … keer wordt ‘gekozen’. De optimalisatie minimaliseert de afwijking van de marges onder de voorwaarde dat alle aantallen geheel zijn. Daardoor wordt voorkomen dat afronding per variabele of per stap (bijv. telkens naar boven of beneden afronden) zich opstapelt en de marges systematisch verstoort.

PTV heeft deze aanpak bewust behouden. Bij de introductie van PopulationSim werd dit algoritme extern gevalideerd en bleek het, mits marges consistent zijn, vergelijkbaar te presteren met andere fittingmethoden. PTV benadrukt dat de belangrijkste problemen in de praktijk zelden in het algoritme liggen: inconsistenties in marges, suppressie[4], ontbrekende cellen en conflicten tussen databronnen bepalen doorgaans de grenzen van wat reproduceerbaar is. Daarom ziet PTV de methode als een "stabiele constante" en investeert het vooral in usability, validatie en onderhoudbaarheid.

Data en schaalniveau
In Duitsland is Mobilität in Deutschland (MiD) de primaire seed-dataset voor de integrale synthesizer. Deze grote huishoudenssteekproef bevat gedetailleerde kenmerken die via regionale classificaties (bijv. stedelijk / suburbaan / landelijk) geschikt zijn gemaakt voor synthese. Waar steekproefdekking beperkt is, worden extra marges gebruikt om plausibele verdelingen te waarborgen[5].

PTV ervaart dat schaalkeuze minder bepalend is voor de kwaliteit dan data-consistentie. Tussen landelijke, regionale en lokale bronnen doen zich regelmatig conflicten voor, wat naar hun oordeel een belangrijker aandachtspunt is dan het specifieke optimalisatie-algoritme. De synthesizer kan op verschillende geografische niveaus draaien, maar de stabiliteit hangt af van de consistentie en volledigheid van de marges.

Privacyregels spelen eveneens een rol. In Duitsland, vergelijkbaar met Nederland, worden microdata alleen op beveiligde servers aangeboden en worden kleine cellen onderdrukt. Dit bemoeilijkt gedetailleerde validatie en leidt soms tot discrepanties tussen totalen. PTV merkt op dat dit geen methodebegrenzing is, maar een structurele eigenschap van de data-omgeving.

Validatie en ervaringen in de praktijk
De nieuwe geïntegreerde synthesizer is nieuw, maar PTV verwacht brede adoptie. Het grootste voordeel is dat gebruikers geen externe software meer hoeven te installeren, een belangrijke barrière die toepassing van PopulationSim jarenlang heeft beperkt.

Validatie richt zich in de eerste plaats op marges en verdelingen. Volgens PTV bepaalt de kwaliteit van de margedata vrijwel volledig de stabiliteit van de resultaten. Waar marges ontbreken of inconsistent zijn, worden pragmatische heuristieken gebruikt. Voor toepassingen voert PTV soms berekeningen uit met niet-geïntegeriseerde populaties, waarbij aantallen niet hoeven te worden afgerond op gehele personen (dus met 'fractionele' personen in de tussenstap), die weken kunnen duren maar dienen als benchmark voor schaalbaarheid. Dit betreft een technische referentie-uitkomst voor vergelijking, niet een populatie die als eindproduct wordt gebruikt.

---

[4] Met suppressie wordt bedoeld dat statistiekleveranciers (zoals CBS) cellen in tabellen (of detailuitsplitsingen) weglaten of afschermen wanneer aantallen te klein zijn, om herleidbaarheid te voorkomen. In de praktijk leidt dit tot 'gaten' of samenvoegingen in marges, waardoor kruistabellen niet sluitend of niet consistent zijn en extra aannames of imputaties nodig worden.

[N] Naast de marges die direct uit MiD volgen, worden aanvullende randtotalen uit andere bronnen opgelegd, bijvoorbeeld voor leeftijdsopbouw, huishoudgrootte, autobezit of inkomensklassen op het beoogde schaalniveau. Deze extra marges sturen de weging en integerisatie zodat de synthetische populatie aansluit op bekende totalen, ook als de steekproef lokaal te klein is of scheef verdeeld.

Beperkingen en aandachtspunten
Het geïntegreerde systeem biedt robuustheid en reproduceerbaarheid, maar heeft ook beperkingen:

- De methodologie is ingebed in commerciële software, waardoor directe toegang tot de optimalisatieprocedure en integerisatie beperkt is.
- De flexibiliteit om alternatieve algoritmen te testen is gering.
- De kwaliteit is sterk afhankelijk van margedata; suppressie, conflicten tussen databronnen en ontbrekende cellen vormen de grootste bron van instabiliteit.
- Privacyregels maken gedetailleerde microvalidatie moeilijk.

PTV benadrukt dat deze beperkingen in de data-infrastructuur vaak belangrijker zijn dan beperkingen in het algoritme zelf.

Relevantie voor SIVMO
Voor SIVMO is de geïntegreerde Visum-module bruikbaar wanneer marges volledig en onderling consistent zijn. De kracht van het systeem ligt vooral in de reproduceerbaarheid en praktische inzetbaarheid binnen gevestigde modelketens. Voor landelijke toepassing is relevant om te weten dat de software en updates afhankelijk zijn van een externe leverancier (PTV) en dat transparantie over rekenstappen beperkt kan zijn.

PTV positioneert de synthesizer als een standaardstap in modelbouw. Dit sluit goed aan bij SIVMO's behoefte aan reproduceerbare processen, maar minder bij situaties waarin volledige openheid van code of methodologische doorontwikkeling vereist is.

Advies van PTV
PTV geeft vier adviezen die direct voortkomen uit hun praktijkervaring:

1 Investeer in datakwaliteit en margedocumentatie. De methode werkt goed wanneer marges betrouwbaar en consistent zijn; dataproblemen zijn bijna altijd de bepalende factor.
2 Beheer en governance zijn cruciaal. Stabiele toepassing vereist duidelijke afspraken over versies, documentatie en kwaliteitscontroles.
3 Houd het syntheseproces eenvoudig en reproduceerbaar. Beperk vrijheidsgraden en voorkom handmatige stappen die tot variatie in uitkomsten leiden.
4 Wees voorzichtig met PopulationSim als open-source alternatief.
   – PopulationSim blijft theoretisch sterk, maar kent structurele onderhoudsproblemen (niet in detail gespecificeerd).
   – Gebruikers lopen risico op gefragmenteerde versies en instabiliteit.
   – Een centrale beheerder is noodzakelijk voor kwaliteit en versiebeheersing.
   – PTV's overstap naar een geïntegreerd systeem was mede ingegeven door de structurele instabiliteit van de Python-implementatie van PopulationSim.

4.3 Department for Transport (DfT, Verenigd Koninkrijk)

Achtergrond en rol
Het DfT is verantwoordelijk voor de nationale transportmodellen in het Verenigd Koninkrijk. Binnen deze context ontwikkelt DfT een nieuwe populatiesynthese-pipeline die voortbouwt op het bestaande National Trip End Model (NTEM). Waar NTEM traditioneel een aggregate trip-end model is, wil DfT een hybride systeem realiseren dat zowel aggregate trip generation ondersteunt als een complete synthetische populatie op persoonsniveau oplevert. De nieuwe aanpak moet de bestaande modelgebruikers blijven bedienen en tegelijk een fundament vormen voor toekomstige activity-based en agent-based modellen.

De synthetische populatie zal volledig consistent zijn met officiële aannames over demografie, huishoudens, werkgelegenheid en regionale ontwikkelingen. Zij moet niet alleen de basis vormen voor trip generation, maar ook als zelfstandig product kunnen worden gepubliceerd. Daarmee beoogt DfT een nationale standaard te creëren die door lokale modellen, consultants en academische partijen kan worden gebruikt.

Methode
De pipeline combineert microdata uit de National Travel Survey (NTS) met officiële prognoses van de Office for National Statistics (ONS). De microdata vormen het zaadbestand; de ONS-projecties bepalen de groei van huishoudens en personen per regio. De synthese wordt uitgevoerd via een iteratief proces waarin sampling, matching en margedoelen worden gecombineerd.

Onder de motorkap gebruikt DfT PopSim als centrale synthesemodule. PopSim wordt aangestuurd in afzonderlijke runs voor verschillende regio's, waarna alle deelpopulaties worden samengevoegd tot één nationale dataset. De keuze voor PopSim was bewust: alternatieven zoals SPENSER bleken minder geschikt voor huishoudstructuren of minder transparant. Voor DfT was het belangrijk om een bestaande, open methode te gebruiken zodat middelen konden worden gebruikt voor organisatie, kwaliteitsborging en documentatie in plaats van het ontwikkelen van een eigen algoritme.

Data en privacyvoorwaarden
De pipeline gebruikt een niet-confidentiële variant van de NTS ("medium restricted"), zodat partners deze kunnen toepassen zonder zware administratieve procedures. Voor het basisjaar wordt gewerkt met 2021-gegevens, ondanks COVID-verstoringen, omdat dit het meest recente en intern consistente volledige demografische bestand is. Pandemie jaren in de NTS worden bewust uitgesloten; eerdere en latere jaren worden gecombineerd tot een representatieve steekproef.

DfT vermijdt waar mogelijk vertrouwelijke bronnen. Voor bedrijfsgegevens is overgestapt op publiek beschikbare bestanden, ondanks het grovere detailniveau. Deze keuze vermindert juridische beperkingen en maakt bredere publicatie mogelijk. De synthetische populatie wordt uiteindelijk geanonimiseerd, zodat deze vrij kan worden gebruikt voor beleidsanalyse, transportmodellen en digitale tweelingen.

Validatie en kwaliteitsborging
Validatie vindt op meerdere niveaus plaats: margecontrole, regionale vergelijkingen, steekproefanalyses en consistentiecontroles over scenariojaren. De pipeline is zodanig opgezet dat wijzigingen in aannames, data of code altijd worden gevolgd via versiebeheer. GitHub wordt gebruikt voor code-reviews, foutregistratie en transparante documentatie.

Externe peer reviews door academische partners vormt een aanvullend onderdeel van de kwaliteitsborging. DfT ziet publicatie van zowel de data als de software als een belangrijke waarborg voor kwaliteit: openheid leidt tot onafhankelijke toetsing en voorkomt dat aannames verborgen blijven.

Uitvoerings- en onderhoudsvraagstukken
DfT benadrukt dat de grootste uitdaging niet wiskundig maar organisatorisch is. Een nationale pipeline vereist continu onderhoud: datasets worden geactualiseerd, afhankelijkheden wijzigen en beveiligingsstandaarden worden aangescherpt. Onderhoud is daarmee een permanente verplichting. Zonder centrale regie kan de pipeline snel verouderen of inconsistent worden.

Daarnaast moet worden voorkomen dat gebruikers de resultaten als te zeker beschouwen. DfT bereidt daarom aanvullende ‘guidance’ voor om duidelijk te maken wat de data wel en niet representeren.

Relevantie voor SIVMO
De aanpak van DfT laat zien dat een landelijke synthesizer alleen duurzaam functioneert wanneer data-invoer, processen, validatie en beheer centraal worden georganiseerd. De technische keuzes zijn toepasbaar in Nederland, maar de belangrijkste lessen liggen op governance- en infrastructuurniveau:

- werk met niet-confidentiële microdata waar mogelijk;
- publiceer zoveel mogelijk aannames, code en resultaten;
- organiseer beheer als continu proces;
- zorg dat alle stappen traceerbaar zijn en reproduceerbaar kunnen worden uitgevoerd.

De DfT-case toont bovendien dat het mogelijk is om een volledige nationale synthetische populatie openbaar te maken, mits het systeem vanaf het begin rond open data en niet-gevoelige microbronnen wordt ontworpen.

Advies van DfT
DfT adviseert om populatiesynthese te benaderen als een proces, niet als een tool. Een stabiele landelijke synthesizer vraagt om duidelijke afspraken over dataversies, governance, documentatie en kwaliteitscontroles. Rijk gedocumenteerde marges zijn volgens DfT belangrijker dan het specifieke algoritme.

Verder beveelt DfT aan om klein te beginnen: start met een beperkte set kenmerken en breid het systeem pas uit als data, stabiliteit en beheer dit toelaten. Overambitie in een vroeg stadium leidt vaak tot instabiliteit, onderhoudsproblemen en onnodige complexiteit.

Tot slot pleit DfT voor samenwerking en openheid: een breed gedragen community van ontwikkelaars, gebruikers en academische partners is essentieel om kennis te behouden, fouten te vermijden en de synthesizer toekomst-bestendig te houden.

4.4 TNO

Achtergrond en rol
TNO heeft ruime ervaring met populatiesynthese in Nederland. De eerste generatie van hun populatiegenerator werd ongeveer acht jaar geleden ontwikkeld in het kader van Urban Tools Next, als onderdeel van een nieuw activity-based model. De oorspronkelijke implementatie in MATLAB gebruikte CBS-microdata als basis en groeide uit tot een landelijke synthetische populatie op PC4-niveau. Deze populatie wordt periodiek geactualiseerd, meest recent met gegevens uit 2023, en vormt een vaste bouwsteen in verschillende activity-based en beleidsgerichte modeltoepassingen.

Methode
TNO gebruikt een IPF-gebaseerde aanpak waarbij drie basisdimensies, te weten leeftijd, geslacht en herkomst, centraal staan. Extra variabelen worden één voor één toegevoegd, omdat simultane multidimensionale IPF in de praktijk instabiel bleek. Dit leidt ertoe dat relaties tussen kenmerken (zoals inkomen en huishoudgrootte) niet automatisch worden behouden. Daarom past TNO postprocessing toe, bijvoorbeeld door autobezit af te stemmen op rijbewijsbezit binnen huishoudens.

De koppeling tussen personen en huishoudens gebeurt via een heuristische procedure die iteratief controleert of de huishoudverdeling overeenkomt met de marges. Leeftijdsrelaties tussen ouders en kinderen worden hierbij expliciet bewaakt. Voor integerisatie wordt een iteratief algoritme toegepast dat restafwijkingen minimaliseert.

CBS-microdata en schaalbeperkingen
De populatiesynthese wordt volledig uitgevoerd binnen de beveiligde CBS-omgeving. Fijnere niveaus dan PC4 kunnen niet worden geëxporteerd vanwege privacyregels, waaronder suppressie en minimum cellengroottes. Waar categorieën te klein zijn, worden schattingen geïmputeerd. Personen worden binnen PC4 aan gebouwen gekoppeld via BAG-gegevens, zodat de woonlocatie plausibel maar niet herleidbaar is.

Volgens TNO vormen vooral kleine categorieën op lokaal niveau een risico voor stabiliteit en nauwkeurigheid. De betrouwbaarheid hangt sterk af van de kwaliteit van marges, definities en consistentie tussen verschillende CBS-tabellen.

Validatieproces en knelpunten
Validatie vindt plaats op twee niveaus. Op zoneniveau worden marges gecontroleerd ten opzichte van PC4-verdelingen. Landelijk worden correlaties vergeleken met patronen in CBS-microdata. Vanaf 2026 wil TNO binnen de remote access (RA) omgeving ook persoonsniveau validatie uitvoeren.

Belangrijke knelpunten zijn inconsistenties tussen CBS-tabellen, beperkte marges voor kleinere gebieden en het ontbreken van transitiemodellen voor scenariojaren. Bij toekomstscenario's worden relaties uit het basisjaar doorgezet naar nieuwe marges, wat tot onnauwkeurigheden kan leiden wanneer kenmerken sterk veranderen.

Toepassing in modellen
De synthetische populatie wordt gebruikt in meerdere TNO-modellen, waaronder ActivitySim en de New Mobility Modeler. Daarnaast wordt de dataset toegepast in onderzoek naar ruimtelijke ontwikkeling, transport- en energiearmoede, gezonde leefstijl en projecten zoals XCARCITY. Door koppeling aan BAG en andere bronnen is de populatie inzetbaar in domeinen buiten verkeer en vervoer.

Koppeling aan regionale modellen vraagt soms om verdere disaggregatie van PC4 naar kleinere zoneringen. Dit gebeurt op basis van BAG-gebouwen, maar de kwaliteit van deze verdelingen varieert per gebied.

Relevantie voor SIVMO
De TNO-aanpak sluit direct aan op beschikbare Nederlandse microdata en laat zien hoe synthese in de praktijk functioneert binnen beleidsmodellen[6]. Tegelijkertijd zijn er duidelijke beperkingen: synthese kan niet worden uitgevoerd buiten de CBS-omgeving, marktpartijen hebben mogelijk geen toegang tot microdata en scenario-toepassingen vereisen veel handmatige bewerking. Hierdoor zou de aanpak minder geschikt kunnen zijn als generieke standaard voor de SIVMO-partners.

Wel biedt TNO's ervaring inzicht in datakwaliteit, privacyvoorwaarden, koppeling aan modellen en praktische uitvoering, elementen die cruciaal zijn voor elke landelijke synthesizer.

Advies van TNO
TNO adviseert om een landelijke synthesizer te baseren op een stabiele, centraal beheerde data-infrastructuur waarin microdatabewerking en kwaliteitscontroles structureel worden georganiseerd. Samenwerking met het CBS is volgens TNO noodzakelijk om marges, definities en updates te standaardiseren.

Daarnaast wijst TNO op de noodzaak om bijzondere huishoudtypen expliciet te modelleren (zoals studentenhuizen, zorginstellingen en gevangenissen) en om leaseauto's zorgvuldig toe te wijzen, omdat deze in databronnen vaak aan bedrijven zijn gekoppeld.

Tot slot pleit TNO voor een samenwerkingsmodel waarin onderzoeksinstellingen en marktpartijen gezamenlijk ontwikkelen en beheren, vergelijkbaar met de bestaande alliantie voor het Programma Strategische Modelvernieuwing van Rijkswaterstaat.

[6] Hier is de term 'beleidsmodellen' gebruikt omdat TNO de populatiesynthese ook op andere beleidsvelden dan vervoer en verkeer inzet

4.5 Centraal Bureau voor de Statistiek (CBS)

Achtergrond en rol
Het CBS beheert de microdatabronnen die in Nederland de basis vormen voor gedetailleerde representaties van huishoudens en personen. Toegang tot deze gegevens vindt uitsluitend plaats binnen de beveiligde Remote Access-omgeving (RA), onder strikte voorwaarden en met gecontroleerde export van uitsluitend geaggregeerde, niet-herleidbare uitkomsten. Daarmee bepaalt het CBS in hoge mate welk detailniveau en welke methoden bij populatiesynthese toepasbaar zijn.

CBS gaf aan dat populatiesynthese voor hen valt binnen het bredere domein van synthetische data. Dit is beleidsmatig en wetenschappelijk een groeiend onderwerp, maar tegelijkertijd omgeven door maatschappelijke en politieke gevoeligheid. Volgens CBS beweegt het instituut in een spanningsveld: enerzijds stimuleert de vraag naar synthetische data innovatie, anderzijds vereist de CBS-wet en de AVG een terughoudende benadering om privacy te beschermen.

In eerdere projecten, zoals het gebruik van microdata in de ontwikkelfase van FEATHERS, speelde CBS vooral een toezichthoudende en faciliterende rol. Methodologische keuzes lagen toen bij externe partijen; CBS borgde veilig datagebruik en bewaakte exportregels. Die rolverdeling is volgens CBS ook voor toekomstige populatiesynthese relevant.

Privacykaders en voorwaarden voor microdata-gebruik
CBS werkt binnen een strikt juridisch kader dat uitgaat van risicobeperking. Microdata mogen uitsluitend binnen RA worden gebruikt. Buiten RA geldt een harde grens: datasets mogen niet fijner zijn dan postcode-4 en moeten sinds 2024 verplicht zijn voorzien van PRAM-ruis. Hierdoor worden waarden vervaagd, categorieën geaggregeerd en wordt het risico op herleidbaarheid verkleind.

Binnen RA is PC6 beschikbaar, maar dit detailniveau mag de omgeving nooit verlaten. Synthese op PC6 is dus alleen mogelijk wanneer de gehele berekening binnen RA plaatsvindt en uitsluitend door gemachtigde instellingen.

CBS wees erop dat risico-inschatting de afgelopen jaren strenger is geworden door technologische ontwikkelingen zoals ‘model inversion attacks’ en ‘membership inference attacks’[7], die theoretisch kunnen leiden tot reconstructie of herkenning van individuen in synthetische data. Hierdoor worden synthetische populaties in eerste instantie behandeld als potentieel privacygevoelig, tenzij aannemelijk is gemaakt dat risico’s voldoende zijn gemitigeerd.

[7] Een **model inversion attack** is een techniek waarbij een buitenstaander probeert om, op basis van de output van een model, kenmerken van individuele personen in de onderliggende dataset te reconstrueren. Daarbij kan gevoelige informatie worden afgeleid zonder directe toegang tot de oorspronkelijke microdata.
Een **membership inference attack** is een methode waarbij een aanvaller probeert te bepalen of een specifiek individu in de trainingsdata van een model voorkomt. Dit kan leiden tot herleidbaarheid, zelfs bij niet-openbare datasets.

Schaalniveau, suppressie en imputatie
Suppressieregels maken dat variabelen met lage aantallen niet beschikbaar worden gesteld of moeten worden geaggregeerd. Dit heeft direct gevolgen voor populatiesynthese: marges op fijnmazige ruimtelijke niveaus zijn slechts beperkt inzetbaar en imputatie van ontbrekende cellen leidt tot extra onzekerheid. CBS onderstreepte dat marges altijd zorgvuldig moeten worden vastgesteld en dat variabelen met kleine aantallen kritisch moeten worden beoordeeld.

Omdat microdatabestanden afkomstig zijn uit verschillende registratiesystemen met uiteenlopende definities en actualisaties, is harmonisatie essentieel. Populatiesynthese is daarom niet alleen een methodologische, maar vooral een datamanagementopgave.

Technische en organisatorische implicaties voor populatiesynthese
CBS benadrukte dat populatiesynthese met microdata alleen veilig en verantwoord mogelijk is binnen RA. Externe synthese op basis van gedetailleerde microdata is niet toegestaan, ongeacht de methode. Buiten RA kan alleen worden gewerkt met PRAM-verrijkte PC4-bestanden of met geaggregeerde output die uit RA is vrijgegeven.

Het genereren van een volledige populatie buiten RA is alleen haalbaar wanneer het proces uitsluitend gebruikmaakt van niet-gevoelige, openbaar toegankelijke data. Zodra microdata of herleidbare patronen nodig zijn, moet synthese worden uitgevoerd door een instelling met RA-machtiging.

Daarnaast werkt CBS aan richtlijnen voor kwaliteitsborging van synthetische gegevens, om zowel statistische kwaliteit als privacybescherming te kunnen beoordelen. Deze richtlijnen worden ontwikkeld in samenwerking met TNO, DUO, UWV, Belastingdienst en universiteiten.

Randvoorwaarden voor SIVMO-toepassing
CBS noemde drie structurele randvoorwaarden voor landelijke populatiesynthese:
- *Schaalniveau:* PC4 is buiten RA het hoogst haalbare detailniveau[8]. Synthese op PC6 kan alleen binnen RA en alleen door gemachtigde instellingen.
- *Consistentie van marges:* marges moeten aansluiten op CBS-publicaties en kunnen niet zonder meer worden gecombineerd uit verschillende externe bronnen. Harmonisatie is noodzakelijk.
- *Dataverwerking binnen een gecontroleerde infrastructuur:* verspreide uitvoering van synthesestappen leidt tot inconsistentie en verhoogde privacyrisico's.

CBS gaf aan dat SIVMO zelf geen microdata kan aanvragen omdat het geen onderzoeksinstelling met rechtspersoonlijkheid is. Een werkbaar model is dat één of meerdere gemachtigde instellingen (zoals TNO of universiteiten) binnen RA de synthese uitvoeren en geaggregeerde bestanden periodiek beschikbaar stellen aan SIVMO-partners.

[8] Dit betreft zowel het ruimtelijk detailniveau waarop woonlocatie mag worden onderscheiden (toewijzing van huishoudens/personen aan gebieden), als het detailniveau van bijbehorende tabellen met onderliggende segmentaties. Buiten RA moeten uitsplitsingen zodanig geaggregeerd blijven dat herleidbaarheid wordt voorkomen.

Daarnaast benadrukte CBS dat synthetische bestanden die buiten RA worden verspreid nooit de indruk mogen wekken dat zij microdata zijn.

Advies van het CBS
CBS adviseert om populatiesynthese te benaderen als een data-intensief proces waarin governance centraal staat. Voor een landelijke oplossing is nodig:

- vaste afspraken over dataversies, definities en margedocumentatie,
- een duidelijk en controleerbaar proces voor dataverwerking binnen RA,
- periodieke kwaliteitscontroles en heldere communicatie over beperkingen en onzekerheden.

CBS adviseert terughoudendheid bij scenario's waarin variabelen geen consistente toekomstprojecties kennen; zonder onderbouwde aannames kunnen synthetische populaties misleidend zijn.

Tot slot onderstreepte CBS dat maatschappelijke gevoeligheid rond synthetische data toeneemt. Transparantie over aannames, methoden en beperkingen is essentieel om vertrouwen te behouden. Een landelijk kader, inclusief toetsingscriteria voor statistische kwaliteit en privacybescherming, wordt door CBS gezien als voorwaarde voor verantwoord gebruik binnen SIVMO.

4.6 Goudappel

Achtergrond en rol
Goudappel ontwikkelt en onderhoudt Octavius, een populatiesynthesemodule die is voortgekomen uit de eerdere Population Synthesizer. De tool is ontworpen voor operationele toepassing in regionale en stedelijke verkeersmodellen en vormt een geïntegreerd onderdeel van een bredere modelketen binnen Omnitrans. Octavius is modulair opgebouwd en kan ook zelfstandig worden ingezet of worden gekoppeld aan andere modelplatformen. De synthesizer is niet domeinspecifiek, maar wordt in de praktijk vooral toegepast binnen verkeer en vervoer.

Methode
Octavius volgt een tweedeling die onderscheid maakt tussen fitting en allocation. Fitting richt zich op het laten aansluiten van huishoud- en persoonskenmerken op marginale verdelingen; allocation op het samenstellen van concrete huishoudens en personen met onderlinge consistentie. Deze scheiding is bepalend voor de ontwerp-filosofie van de tool en biedt flexibiliteit in methoden en scenario's.

De huidige versie van Octavius gebruikt een deterministische aanpak zonder sampling. Waar eerdere generaties IPF en IPU toepasten, bleek dit methodologisch kwetsbaar bij sterke afhankelijkheden tussen huishouden en persoon. Daarom is gekozen voor een geïntegreerde fitting methode op basis van non-negative least squares (NNLS)[9], waarin huishoud- en persoonskenmerken gelijktijdig worden gefit. De continue

[9] NNLS is een optimalisatiemethode waarmee een continue verdeling wordt geschat die zo goed mogelijk aansluit op opgelegde marges, onder de voorwaarde dat alle waarden niet-negatief zijn. De techniek wordt gebruikt om een realistische, maar nog niet-geintegeriseerde populatieverdeling te bepalen.

oplossing wordt daarna via de Statistical Noise Elimination Technique (SNET)[10] gediscretiseerd, zodat een stabiele en reproduceerbare populatie ontstaat zonder Monte Carlo-ruis. Dit maakt Octavius geschikt voor toepassingen waarin scenariovergelijking en reproduceerbaarheid essentieel zijn.

Data-inrichting en randvoorwaarden
Octavius is niet gebonden aan specifieke databronnen, maar werkt in de praktijk met CBS-buurtstatistieken, ODIN, MPN en aanvullende projectafhankelijke marges. Omdat CBS-microdata niet buiten de RA-omgeving mogen worden verwerkt, moeten relaties tussen kenmerken worden gereconstrueerd uit gecombineerde bronnen. Dit gebeurt op basis van geaggregeerde bronnen en (waar beschikbaar) vrijgegeven statistieken uit RA; buiten RA worden geen CBS-microdata gebruikt. Daarmee wordt niet geprobeerd microdata te reproduceren, maar worden verbanden benaderd binnen de geldende randvoorwaarden rond detailniveau en herleidbaarheid. Daarbij kunnen afhankelijkheden worden soms geïmputeerd met behulp van oudere, grotere steekproeven zoals OVG of MON. De stabiliteit van huishoudstructuren in Nederland maakt dergelijke reconstructies in veel gevallen bruikbaar.

De synthesizer kan op verschillende schaalniveaus worden toegepast, maar de kwaliteit van de populatie hangt sterk af van de volledigheid en consistentie van de marges. Synthese op een niveau fijner dan PC4 is alleen mogelijk wanneer marges voldoende stabiel zijn geaggregeerd. Dit sluit aan bij het gegeven dat detailniveau buiten RA beperkt is en dat fijnmazige kruistabellen snel tot suppressie of aggregatie leiden. Voor sommige kenmerken worden uitsluitingen of harde randvoorwaarden ingesteld om onmogelijke combinaties te voorkomen.

Validatie en kwaliteit
Octavius voert interne margevalidatie uit op zowel huishoud- als persoonsniveau. Dit vormt de primaire controle, omdat in veel projecten externe microdata niet beschikbaar zijn of niet gebruikt mogen worden. Verdere validatie gebeurt via kruistabellen, gedragsvergelijkingen of expertbeoordeling, maar wordt in de praktijk beperkt uitgevoerd vanwege beschikbare middelen en opdrachtgeverseisen.

De deterministische opzet leidt tot robuuste en consistente resultaten. Foutmarges ontstaan vooral bij zeldzame combinaties of lege cellen. In zulke gevallen worden categorieën samengevoegd of uitgesloten, met doorgaans beperkte invloed op de totale populatie. De stabiliteit maakt Octavius geschikt voor scenarioanalyses over meerdere jaren of backcasts.

Gebruik en toepasbaarheid in modellen
Octavius wordt ingezet als generieke synthesemodule binnen uiteenlopende transportmodellen, waaronder trip-based, tour-based modellen, en in sommige gevallen als bouwsteen richting activity-based toepassingen. De tool genereert de basispopulatie waar andere modules, zoals voertuigbezit, activiteitenschema's of tourvorming, op voortbouwen. De modulariteit maakt koppeling met externe variabelen en modellen mogelijk. Hoewel de synthesizer is ontwikkeld voor toepassing

[10] SNET is een deterministische integerisatiemethode die een continue oplossing omzet in discrete huishoudens en personen, zonder gebruik te maken van toevalsprocessen. Hierdoor ontstaat een reproduceerbare populatie zonder Monte Carlo-ruis.

in transportmodellen, kan het ook worden ingezet in domeinen zoals ruimtelijke ordening, gezondheidszorg of onderwijsplanning.

Sterke en zwakke punten
Sterke punten zijn de deterministische werking, de reproduceerbaarheid, de flexibiliteit in databronnen en randvoorwaarden, en de modulaire integratie met bredere modelketens. Octavius is geschikt voor operationeel gebruik in regionale contexten en middelgrote datasets.

Beperkingen ontstaan wanneer zeer gedetailleerde of microdata gestuurde validatie nodig is. Ook is de methode afhankelijk van zorgvuldig samengestelde marges; onvolledige of inconsistente tabellen verminderen de kwaliteit van de uitkomsten. Randvoorwaarden moeten handmatig worden geconfigureerd, wat expertise vereist.

Relevantie voor SIVMO
De aanpak van Goudappel illustreert dat populatiesynthese een modulaire keten is en geen enkelvoudige methode. De scheiding tussen fitting en allocation, de deterministische aanpak en de aandacht voor integerisatie zijn voor SIVMO waardevolle ontwerpprincipes. Tegelijkertijd laat de praktijk zien dat regionale marges en gebrek aan microdatatoegang beperkingen opleggen aan landelijke uniformiteit.

De synthesizer zelf is niet open source. Dat maakt gezamenlijke ontwikkeling en gezamenlijk onderhoud door meerdere SIVMO-partners minder vanzelfsprekend, en kan leiden tot afhankelijkheid van één leverancier en een kennisvoorsprong buiten het partnerschap. Een landelijke standaard kan ook worden ingericht als een gedeelde codebasis die door meerdere partijen wordt ontwikkeld en beheerd, met afspraken over governance, kwaliteit, versiebeheer en releasebeleid. Wanneer broncode en onderhoudsrechten niet gedeeld kunnen worden, zijn aanvullende afspraken nodig over documentatie, auditbaarheid, overdraagbaarheid en continuïteit.

In dit onderzoek is niet vastgesteld of en onder welke voorwaarden SIVMO-partners de synthesizer als afzonderlijk product kunnen afnemen, dan wel (mede-)eigenaar kunnen worden; evenmin is vastgesteld of co-ontwikkeling of het delen van (delen van) broncode mogelijk is. Dit vraagt nadere uitwerking met de leverancier.

Advies van Goudappel
Goudappel pleit voor een modulaire, open en herbruikbare basis voor populatiesynthese in Nederland. Het vermijden van dubbele ontwikkeling is belangrijk: bestaande methoden en tools bieden voldoende fundament voor een gedeelde infrastructuur, mits deze transparant wordt ingericht. Voor SIVMO betekent dit dat:

- fitting en allocation als gescheiden modules moeten worden ontworpen;
- deterministische methoden een goede basis vormen voor reproduceerbare scenario's;
- validatie meer moet omvatten dan margecontrole alleen;
- samenwerking en gezamenlijk beheer noodzakelijk zijn om versnippering te voorkomen;
- open ontwikkeling de voorkeur heeft boven geïsoleerde commerciële implementaties.

4.7 Significance

Achtergrond en rol
Significance heeft ruime ervaring met populatiesynthese in Nederland en Vlaanderen. Het bureau ontwikkelt en past zowel synthese- als transitiemodellen toe en reflecteert vanuit die praktijk op methodologische keuzes. Significance gaf aan dat het literatuuranalyse een helder overzicht biedt van synthetische populatiemethoden, met een indeling die goed aansluit bij de internationale literatuur.

Methodologische reflectie: synthese en transitie
Volgens Significance is het essentieel om een expliciet onderscheid te maken tussen populatiesynthese voor een basisjaar en transitiemodellen voor populatieontwikkeling over de tijd. Populatiesynthese genereert een momentopname op basis van waargenomen marges en microdata. Dit is de aangewezen methode voor basisjaren van verkeers- en vervoersmodellen, omdat stabiliteit, controleerbaarheid en aansluiting op statistieken belangrijk zijn.

Transitiemodellen volgen een andere benadering: de populatie evolueert via levensloopgebeurtenissen zoals geboorte, sterfte, migratie, huishoudvorming en loopbanen. Hiermee ontstaat een meer consistente ontwikkeling over de tijd, met behoud van micro-niveaurelaties zoals inkomen, huishoudtype of opleiding. Deze modellen bieden rijkere dynamiek, maar zijn methodologisch en organisatorisch zwaarder en worden in de praktijk minder frequent toegepast.

Ervaringen met transitiemodellen
Vlaanderen kent ervaring met transitiemodellen. Een eerste generatie model integreerde alle demografische processen in één simulatie. Hoewel dit theoretisch sterk was, bleek het onderhoudsintensief en complex. Het model is uiteindelijk verlaten vanwege de hoge ontwikkelkosten en beperkte operationele inzet.

In vervolgjaren is overgestapt op een lichtere aanpak waarin per scenariojaar nieuwe populaties worden gesynthetiseerd met optimalisatietechnieken. Deze methode is beheersbaar maar ontbeert interjaar-consistentie. Hierdoor ontstaan losse momentopnames (“snapshots”) zonder een natuurlijke relatie in de tijd, bijvoorbeeld bij vergrijzing, huishouddynamiek of sociaal-economische ontwikkeling.

Transitiemodellen bieden in dat opzicht een logischer en rijker verloop, maar vragen veel onderhoud en worden vaak slechts incidenteel gebruikt. De inzet ervan is vooral zinvol wanneer jaren-op-jaren consistentie of life-eventdynamiek relevant is.

Toepassing binnen vervoers- en verkeersmodellen
De meeste transportmodellen in Nederland en Vlaanderen werken met vaste projectiejaren zoals 2025, 2035 of 2050. Populaties worden per jaar apart gesynthetiseerd op basis van externe prognoses van CBS, ABF of sectorale bronnen. Deze aanpak is praktisch en sluit aan bij huidige planningsprocessen, maar mist de interne logica van transitiemodellen. Kenmerken zoals vergrijzing, onderwijsloopbanen of migratieachtergrond worden dan indirect meegenomen via marges, niet via micro-evolutie.

Een transitiemodel kan waardevol zijn als externe schil: eerst simulatie van demografische veranderingen, daarna selectie of synthese van scenariojaren, en ten slotte koppeling aan verkeersmodellen via activiteitengeneratie of tourvorming. Deze indirecte benadering combineert dynamiek met beheersbaarheid.

Sterke en zwakke punten van beide benaderingen
Populatiesynthese is sterk in stabiliteit, transparantie en reproduceerbaarheid. De methode is goed geschikt voor basismodellen, agenttypen en verkeersprognoses met beperkte eisen aan microdynamiek.

Transitiemodellen bieden meer interne samenhang en realistische evolutie, maar vergen gespecialiseerde kennis, intensief beheer en institutionele borging. De ervaring in Vlaanderen laat zien dat deze modellen alleen duurzaam bruikbaar zijn wanneer eigenaarschap, onderhoud en intersectorale afstemming structureel geregeld zijn.

Relevantie voor SIVMO
Voor SIVMO bevestigt Significance dat populatiesynthese de logische basis vormt voor het opstellen van een nationale populatie op tijdstip t_0. Voor toekomstscenario's kunnen transitiemodellen relevant zijn, maar alleen wanneer de beleidsvragen vragen om consistentie over meerdere jaren of wanneer life events substantieel bijdragen aan mobiliteitsgedrag. De praktijk leert dat niet elk project dat niveau van detail vereist.

Advies van Significance
Significance adviseert om het onderscheid tussen synthese (basisjaar) en transitie (toekomstontwikkeling) expliciet op te nemen in de methodologische beoordeling. Populatiesynthese vormt de basis, maar transitiemodellen kunnen een aanvullende rol spelen wanneer:

- tijdsdynamiek en levensloopprocessen relevant zijn;
- interjaar-consistentie noodzakelijk is;
- gedragsmodellen of agent-based benaderingen profiteren van rijke microstructuur.

Tegelijkertijd vraagt deze keuze ook om aandacht voor governance: transitiemodellen zijn alleen houdbaar bij helder eigenaarschap, structurele financiering en institutionele inbedding. Voor SIVMO betekent dit dat de technische, organisatorische en beleidsmatige consequenties evenredig moeten worden meegewogen.

4.8 Overkoepelende bevindingen

De interviews laten een duidelijk en breed gedragen beeld zien van populatiesynthese als een proces dat methodologie, data-infrastructuur, governance en toepassing onlosmakelijk met elkaar verbindt. Hoewel de organisaties uiteenlopende rollen vervullen komt in alle gesprekken naar voren dat de effectiviteit van populatiesynthese minder wordt bepaald door het gekozen algoritme en meer door de kwaliteit van data, bestuurlijke inbedding en de organisatie van het totale proces.

1. Methodologische verschillen bestaan, maar bepalen zelden de kwaliteit
De gesprekspartners maken gebruik van uiteenlopende technieken: entropie-maximalisatie (PTV), sampling en matching (DfT), sequentiële IPF met postprocessing (TNO) en deterministische NNLS/SNET-constructies (Goudappel). Significance benadrukt bovendien het onderscheid tussen synthese (t_0) en transitie (ontwikkeling over tijd). Ondanks deze verschillen is de gedeelde conclusie dat methoden in de praktijk vergelijkbaar presteren zolang marges volledig, consistent en stabiel zijn.

Naast het algoritme speelt ook een verschil in 'opzet': sommige synthesizers werken met een steekproef die via weging en integerisatie wordt opgeschaald naar een volledige populatie (sampling plus enumeration), terwijl andere benaderingen leunen op kenmerken uit administratieve registers (zoals leeftijd, geslacht, huishoudtype en inkomen) binnen microdata om per individu een volledige populatie te reconstrueren. Dit onderscheid raakt vooral de beschikbaarheid van microdata en de privacy-randvoorwaarden, minder het fittingprincipe zelf.

Variatie in datakwaliteit, suppressie, conflicten tussen bronnen en beperkt beschikbare microdata heeft een veel grotere invloed op output dan de keuze tussen algoritmen. Dit maakt dat methodologische discussies niet los kunnen worden gezien van data-infrastructuur en governance.

2. Microdatatoegang is bepalend voor wat methodologisch mogelijk is
TNO en CBS illustreren dat hoogwaardige synthese in Nederland sterk afhankelijk is van microdata binnen de Remote Access (RA) omgeving. Buiten RA zijn alleen PRAM-verrijkte PC4-bestanden beschikbaar[11]; dit beperkt gedetailleerde modellering én validatie. PTV en Goudappel ondervinden vergelijkbare beperkingen in Duitsland en Nederland: suppressie, beperkte herleidbaarheid en inconsistente tabellen zetten grenzen aan wat methodologisch haalbaar is.

Voor SIVMO betekent dit dat toegang tot microdata niet gelijk verdeeld is. Landelijke toepassingen moeten daarom rekening houden met een gelaagde datastructuur waarin sommige stappen alleen door gemachtigde instellingen kunnen worden uitgevoerd.

3. Reproduceerbaarheid en onderhoudbaarheid zijn belangrijke randvoorwaarden
Alle organisaties benadrukken dat populatiesynthese structureel onderhoud vraagt. DfT spreekt over "you have to run just to stand still", PTV verving PopulationSim vooral vanwege onderhoudsproblemen, en TNO besteedt een groot deel van de tijd aan datavoorbewerking en harmonisatie. Ook Goudappel wijst op het belang van stabiliteit in de modellen (dezelfde input en instellingen moeten dezelfde uitkomst geven bij reproductie) en Significance benadrukt het belang van institutionele borging bij transitiemodellen: eigenaarschap, financiering en beheer moeten expliciet geregeld zijn.

[11] PRAM (Post Randomisation Method) is een techniek waarbij waarden van variabelen in microdata met een bepaalde kans worden aangepast om herleidbaarheid te verkleinen.

Een robuuste synthesizer vergt dus meer dan een goed algoritme: versiebeheer, governance, documentatie, duidelijk eigenaarschap en een centrale beheerstructuur blijken noodzakelijk in alle contexten.

4. Transparantie en toetsbaarheid zijn essentieel
DfT ziet openheid, publicatie van code en peer review als kern van kwaliteitsborging. PTV biedt reproduceerbaarheid maar minder inzicht in interne optimalisatie vanwege commerciële software. TNO en CBS opereren binnen RA, waar methoden transparant kunnen zijn maar output en validatie beperkt deelbaar zijn. Goudappel werkt met een gesloten workflow, waarbij de mate van documentatie en toetsbaarheid in de praktijk afhankelijk is van interne vastlegging en overdracht naar externe partijen.

In alle gevallen blijkt toetsbaarheid een voorwaarde voor vertrouwen, maar de praktische invulling varieert. Voor een landelijke aanpak betekent dit dat transparantie niet alleen technisch moet worden geborgd, maar ook juridisch en organisatorisch.

5. Margedata vormen overal het structurele knelpunt
PTV, DfT, TNO en Goudappel wijzen op margedata[12] als bepalend voor stabiliteit. Incomplete, inconsistente of onderdrukt gepubliceerde marges leiden direct tot instabiliteit, noodzaak tot imputatie of handmatige heuristieken. CBS bevestigt dat verschillen in definities, suppressie en PRAM-noise de bruikbaarheid van marges beïnvloeden.

De kwaliteit van de populatie volgt dus rechtstreeks uit de kwaliteit, definitiestabiliteit en actualiteit van de marges, niet uit de complexiteit van het algoritme.

6. Toepassingen verschillen, maar vereisen allemaal een stabiele basispopulatie
Alle partijen bevestigen dat een synthetische populatie een basis is voor trip-based, tour-based en activity-based modellen. PTV en Goudappel integreren synthese in hun modelplatforms, DfT gebruikt de populatie als nationale standaard, TNO voedt er diverse beleidsdomeinen mee, en Significance plaatst synthese als basisjaarcomponent in transitie trajecten.

De gedeelde behoefte is een stabiel en uniform uitgangspunt dat herhaald kan worden over jaren, projecten en domeinen.

7. Toekomstscenario's vragen om expliciete keuzes tussen stabiliteit en dynamiek
Significance maakt duidelijk dat synthese voor een basisjaar en transitie voor scenariojaren fundamenteel verschillende processen zijn. TNO worstelt met scenariojaren omdat er geen microdata voor toekomstrelaties bestaan. DfT focust eerst op synthesepopulaties alvorens dynamiek toe te voegen.

[12] Met margedata worden hier de opgelegde randtotalen en verdelingen bedoeld waarop de synthese wordt 'gefitt', zoals aantallen personen of huishoudens per gebied en per categorie (bijv. leeftijdsklasse, huishoudgrootte, inkomen, autobezit)

Er ontstaat consensus dat scenario's gebaseerd moeten zijn op expliciete, transparante en herleidbare aannames, en dat, naast een synthethische populatie voor het basisjaar, een methode nodig is om populaties voor toekomstjaren te construeren. Een jaar-op-jaar transitiemodel is één optie, maar niet de enige: in veel toepassingen volstaan ook scenario-opschaling op basis van bijvoorbeeld marges of trendprojecties. Een transitiemodel is vooral relevant wanneer de beleidsvraag expliciet vraagt om interjaarlijkse samenhang en pad-afhankelijkheid (bijv. consistentie van levensloop- of huishoudtransities over meerdere jaren)..

8. Samenwerking is noodzakelijk om versnippering te voorkomen
Alle partijen benoemen het belang van gezamenlijke ontwikkeling, gedeelde standaarden en centraal beheer. TNO en CBS wijzen op juridische randvoorwaarden. Goudappel en PTV benadrukken dat dubbele ontwikkeling inefficiënt is. DfT laat zien hoe een nationale community via open publicatie kan worden gevormd. Significance wijst op het belang van institutionele borging.

Gezamenlijk ontstaat het beeld dat SIVMO alleen duurzaam kan opereren met een model waarin data, definities, validatie en updates centraal zijn georganiseerd.

Samenvattend, de interviews tonen dat succesvolle populatiesynthese niet primair afhangt van het gekozen algoritme, maar van de kwaliteit en toegankelijkheid van data, duidelijke governance, reproduceerbare processen en een gezamenlijk beheer dat versnippering voorkomt.

Voor SIVMO betekent dit dat een landelijke aanpak alleen kan slagen wanneer data, processen en governance centraal en uniform worden ingericht. Een werkbare synthesizer vraagt om duidelijke afspraken, consistente definities en periodieke kwaliteitscontroles. Tegelijkertijd moet rekening worden gehouden met structurele randvoorwaarden, zoals de beperkingen rond microdata toegang en het schaalniveau waarop synthese verantwoord kan plaatsvinden. De inzichten uit dit hoofdstuk vormen daarmee de basis voor het opstellen van een methodekeuze, architectuur en werkproces voor de Nederlandse context.

5
STRATEGY
SOLUTION
SUCCESS
OPTIMIZE
ACHIEVEMENT

5 Beoordeling

Dit hoofdstuk beoordeelt de onderzochte methoden en tools op sterke en zwakke kanten, met het oog op toepassing in de Nederlandse modelpraktijk. De interviews laten zien dat uitkomsten vaak meer worden bepaald door datakwaliteit, definities en beheer dan door het algoritme alleen. De beoordeling plaatst methoden daarom in hun data- en governancecontext en levert geen ranglijst, maar een set afwegingen en randvoorwaarden per benadering. De analyse is gebaseerd op literatuur, gebruikerservaringen en toolbeschrijvingen.

5.1 Beschouwing van methoden

De methodologische benaderingen voor populatiesynthese verschillen sterk in structuur, aannames en vereisten aan data. De vijf onderscheiden families – iteratieve ophoogmethoden, optimalisatie-gebaseerde methoden, probabilistische reconstructie, simulatieve/generatieve modellen en datafusie/hybride technieken – kennen elk hun eigen toepassingsgebied, voordelen en beperkingen. In deze paragraaf worden deze families afzonderlijk beschouwd, los van specifieke tools.

Iteratieve ophoogmethoden zoals IPF en IPU bieden een eenvoudige, transparante en deterministische aanpak. Ze zijn goed inzetbaar bij voldoende microdata en betrouwbare randtotalen, en worden vaak toegepast voor het opstellen van populaties voor trip- en tour-based modellen. De reproduceerbaarheid en het lage rekenkundige gewicht maken ze aantrekkelijk in beleidspraktijk. Tegelijk hebben deze methoden moeite met minder gangbare populatie groepen, en modelleren ze geen correlatiestructuren, tenzij deze al aanwezig zijn in de steekproef. Daarmee zijn ze minder geschikt voor situaties met hoge dimensionaliteit (veel kenmerken en categorieën tegelijk).

Optimalisatie-gebaseerde methoden zoals entropiemaximalisatie of minimalisatie van KL-divergentie bieden meer flexibiliteit. Ze maken het mogelijk om meerdere constraints tegelijk te hanteren en leveren robuustere resultaten bij lege of onvolledige celcombinaties. Deze methoden zijn vooral relevant bij complexere modellen, zoals activity-based modellen, of wanneer marges op meerdere niveaus moeten worden gecombineerd. Nadeel is de hogere complexiteit van implementatie, en de noodzaak tot nauwkeurige parameterinstellingen. Desondanks kunnen ze een potentiële basis vormen voor standaardisatie, maar toepassing in Nederland vraagt nog om verdere uitwerking, kennisopbouw en beheerafspraken.

Probabilistische reconstructiemethoden richten zich op het genereren van populaties door sampling uit een steekproefbestand. Zij bieden flexibiliteit bij beperkte of onvolledige data, en maken onzekerheidsmodellering mogelijk. Dit is vooral nuttig in verkennende analyses en agent-based modellen. Nadelig is dat de uitkomsten per run kunnen variëren, waardoor deze methoden minder geschikt zijn voor toepassingen waarin consistentie en reproduceerbaarheid vereist zijn, tenzij 'seeds' worden gebruikt. Afhankelijk van de gekozen post-processing kan aanvullende correctie nodig zijn om marges exact te laten aansluiten.

Simulatieve en generatieve modellen bouwen populaties op vanuit gedragsregels of simulaties van demografische ontwikkeling. Ze zijn vooral geschikt voor toekomstgerichte toepassingen, of modellen waarin gedrag, context en populatie-evolutie samenhangen. De keerzijde is een hogere modelcomplexiteit en afhankelijkheid van aannames die lastig te valideren zijn. In Nederland zijn er toepassingen (zoals SPARK), maar de inzet vraagt doorgaans expliciete aannames, aanvullende validatie en beheer om consistent gebruik in een modelketen te borgen.

Datafusie en hybride technieken, inclusief machine learning en deep generative methods zoals VAEs en GANs, bieden interessante perspectieven bij incomplete of privacygevoelige data. Ze maken het mogelijk om datasets te koppelen, ontbrekende kenmerken te imputeren, of geheel synthetische populaties te genereren. Hun kracht ligt vooral in experimentele of onderzoeksgerichte omgevingen. Voor structurele toepassing in transportmodellen is verdere ontwikkeling en validatie nodig.

De analyse laat zien dat geen enkele methode universeel toepasbaar is. De keuze hangt af van de context: modeltype, schaal, data, en eisen aan transparantie en robuustheid. Voor de SIVMO partners lijkt een deterministische, optimalisatie-gebaseerde benadering voorlopig de meest geschikte basis, eventueel aangevuld met probabilistische of simulatieve elementen in specifieke toepassingen. Meer flexibele methoden kunnen aanvullend waardevol zijn in specifieke domeinen, mits voldoende gevalideerd.

5.2 Beoordelingskader voor de tools

Voor de beoordeling zijn zes kerncriteria gehanteerd:

1. *Methodologische geschiktheid*. Is de methode geschikt voor het type model waarin de populatie wordt toegepast (trip, tour, AcBM)? Hoe gaat de methode om met complexiteit en margestructuur?
2. *Toepasbaarheid en schaal.* In hoeverre is de tool inzetbaar op verschillende geografische niveaus (landelijk, regionaal, lokaal) en populatiegroottes?
3. *Reproduceerbaarheid en robuustheid.* Levert de methode stabiele uitkomsten bij gelijke invoer? Hoe gevoelig is het systeem voor lege cellen of kleine populatiegroepen?
4. *Validatie in de praktijk.* Is er sprake van toetsing aan waarnemingen of benchmarks? Zijn validatieroutines ingebouwd of beschikbaar?
5. *Beheer en onderhoud.* Is de tool goed gedocumenteerd, configureerbaar en onderhoudbaar? Is ondersteuning beschikbaar?
6. *Toegankelijkheid en openheid.* Is de tool open source, beschikbaar voor derden, en aanpasbaar aan de Nederlandse context?

Deze criteria zijn toegepast op een selectie van acht tools die representatief zijn voor de verschillende methodologische families: Quad, SigPopu, Octavius, PopulationSim, PopGen, PopSynWin, MATSim en SynthPop. De resultaten zijn samengevat in de volgende paragraaf.

5.3 Samenvattende vergelijking van de tools

De populatiesynthese tools verschillen in methode, schaalbaarheid, reproduceer-baarheid en toegankelijkheid. Tabel 1 geeft een overzicht van acht representatieve tools die in de Nederlandse context relevant zijn. Zij vertegenwoordigen de belangrijkste methodologische families en zijn geselecteerd op basis van bekendheid, documentatie en gebruik in onderzoeks- of beleidspraktijk. De tabel geeft per tool aan hoe deze scoort op zes criteria: methodologische geschiktheid, toepasbaarheid en schaal, reproduceerbaarheid, validatie, beheer en onderhoud, en toegankelijkheid.

Tabel 1 Beoordeling van synthesizers op zes kerncriteria

Tool	Methode	Modeltype	Reproductie	Openbaar	Op te schalen	Validatie-ervaring	Opmerkingen
Quad	Newton–Raphson	Trip / Tour	✓	✕	✓	✓	Kwetsbaar bij kleine of lege cellen
SigPopu	KL-optimalisatie	Trip / Tour	✓	✕	✓	✓	Niet open source; In beheer bij RWS, robuust resultaat
Octavius	NNLS + SNET	Trip / Tour / AcBM	✓	✕	✓	✓	Vereist goed samengestelde marges
PopulationSim	Entropie-maximalisatie + integerisatie	Trip / Tour / AcBM	✓	✓	✓	✓	Open source, gevoelig voor inconsistenties in marges
PopGen	IPU + sampling	Trip / Tour	~	✓	✓	✓	Flexibel, ondersteunt onzekerheids-analyse
PopSynWin	IPF	Trip / Tour	✓	✓	✕	~	Geschikt voor onderwijs en kleinschalige projecten
MATSim	Agent-based simulatie	AcBM	✓	✓	✓	✕	Populatie-opbouw vereist aanvullende scripts
SynthPop	CART / regressiemodellen	Verkennend / research	✓	✓	~	✓	Gericht op privacy en synthetische microdata

Bron: Analyse Panteia. Legenda: ✓ = sterk punt, ~ = neutraal of contextafhankelijk, ✕ = beperking

De vergelijking laat drie patronen zien. Ten eerste blijken **deterministische optimalisatie-methoden**, zoals SigPopu en PopulationSim, geschikt voor situaties waarin reproduceerbaarheid en stabiliteit centraal staan. PopulationSim scoort hoog op flexibiliteit en openheid. SigPopu is sterk in robuustheid bij complexe marges, maar is niet open source. De tool is in beheer bij RWS en daarmee in de Nederlandse praktijk beschikbaar voor brede toepassing, vergelijkbaar met QUAD.

Ten tweede tonen tools als **Octavius** en **PopGen** dat hybride en probabilistische benaderingen goed functioneren wanneer marges afkomstig zijn uit meerdere bronnen of wanneer onzekerheid moet worden meegenomen. Deze tools zijn waardevol in regionale analyses of scenarioverkenningen. De kwaliteit hangt wel af van de invoerdata.

Ten derde geldt dat tools met een bredere functionele scope, zoals **MATSim** en **SynthPop**, gericht zijn op specifieke onderzoeksdomeinen. Zij zijn minder geschikt als generieke standaard voor landelijke populatiesynthese, maar wel waardevol binnen gespecialiseerde toepassingen of methodologische experimenten.

De tabel laat zien dat de keuze afhangt van de toepassing, het schaalniveau, de eisen aan reproduceerbaarheid en de mate waarin transparantie en toegankelijkheid vereist zijn.

5.4 Relevantie voor de Nederlandse context

De vergelijking van de tools laat een aantal observaties zien die relevant zijn voor toepassing binnen de Nederlandse modelpraktijk en specifiek voor SIVMO.

Een eerste observatie is dat **deterministische, optimalisatie-gebaseerde tools** in de praktijk het meest consistent presteren. Tools zoals SigPopu, PopulationSim en Octavius leveren stabiele en reproduceerbare populaties, mits de marges intern consistent zijn en goed zijn gedocumenteerd. Dit sluit aan bij de ervaringen van PTV en Goudappel, die benadrukken dat de kwaliteit van de invoerdata doorgaans bepalender is dan het gekozen algoritme. Voor beleidsmodellen, waarin herhaal-baarheid en uitlegbaarheid essentieel zijn, vormen deze tools daarom een logische basis.

Een tweede observatie is dat **openheid en onderhoudbaarheid sterk uiteenlopen**. Open tools zoals PopulationSim en PopGen bieden transparantie en aanpasbaarheid, maar vragen ook om actieve regie op versiebeheer, documentatie en configuratie. Gesloten tools zoals SigPopu en Octavius zijn in de regel stabieler in gebruik, maar beperken de mogelijkheid tot methodologische doorontwikkeling en externe controle. De interviews maken duidelijk dat deze afweging niet primair technisch is, maar samenhangt met governance: wie beheert de tool, wie is verantwoordelijk voor updates en hoe wordt kennis geborgd.

Een derde observatie betreft **probabilistische en sampling-gebaseerde tools**. Deze bieden meer flexibiliteit bij onvolledige of inconsistente data en maken het mogelijk om onzekerheid expliciet te analyseren. De DfT-case laat zien dat sampling ook in een nationale context toepasbaar is, mits reproduceerbaarheid wordt geborgd via vaste

seeds en een strak vastgelegde werkwijze. Tegelijk blijft variatie tussen runs een aandachtspunt. Voor toepassingen waarin vaste, exact herhaalbare uitkomsten vereist zijn, zijn deze tools daarom minder geschikt als standaardoplossing, maar wel waardevol als aanvullend instrument.

Een vierde observatie is dat **hybride tools goed kunnen functioneren**, maar moeilijk opschaalbaar zijn naar een uniforme landelijke standaard. Octavius laat zien dat het combineren van meerdere databronnen praktisch haalbaar is wanneer microdata niet toegankelijk zijn. De kwaliteit van de uitkomsten is dan echter wel afhankelijk van de gemaakte aannames en de stabiliteit van gereconstrueerde relaties. Dit maakt dergelijke tools geschikt voor specifieke toepassingen.

Een vijfde observatie is dat **validatie in de praktijk vaak beperkt blijft tot marge-controles**. Slechts enkele tools beschikken over expliciete routines voor bredere consistentietoetsen, zoals kruistabellen of vergelijkingen met externe referenties. De interviews bevestigen dat diepgaande validatie meestal wordt beperkt door tijd, middelen of datatoegang. Dit betekent dat verschillen tussen tools in validatie-mogelijkheden in de praktijk kleiner zijn dan theoretisch wordt verondersteld.

Tot slot blijkt dat **beheer en organisatie een doorslaggevende rol spelen** in de geschiktheid van tools. Tools die technisch sterk zijn, maar waarvoor geen structureel beheer, documentatie en ondersteuning is ingericht, zijn op termijn kwetsbaar. Dit geldt met name voor open-source oplossingen zonder duidelijke eigenaar, maar ook voor gesloten tools wanneer kennis sterk bij één leverancier of team is geconcen-treerd. De ervaringen van PTV, DfT en TNO wijzen erop dat een tool alleen duurzaam inzetbaar is wanneer beheer expliciet is belegd en onderdeel vormt van het totale syntheseproces.

Samenvattend laten de observaties zien dat de geschiktheid van populatiesynthese tools minder wordt bepaald door hun theoretische methode en meer door hun praktische inpasbaarheid: datakwaliteit, reproduceerbaarheid, beheer en governance. Deze inzichten vormen de basis voor de beoordeling van de tools in de Nederlandse context in de volgende paragrafen.

5.5 Tussenconclusie

De analyse van methoden toont aan dat er geen eenduidige ‘beste’ aanpak bestaat. De geschiktheid van een populatiesynthese methode hangt sterk af van de context waarin deze wordt toegepast. In de Nederlandse praktijk, waarin uiteenlopende modeltypen naast elkaar bestaan, is behoefte aan een flexibele, maar robuuste basisbenadering.

Optimalisatie-gebaseerde methoden bieden daarvoor op dit moment het beste perspectief. Zij combineren stabiliteit, schaalbaarheid, reproduceerbaarheid en aanpasbaarheid aan verschillende geografische niveaus en randvoorwaarden. Zowel SigPopu als PopulationSim scoren hier goed, met als kanttekening dat SigPopu gesloten is en PopulationSim nog verdere aanpassing aan Nederlandse data vereist.

Andere methoden, zoals iteratieve ophoogtechnieken (IPF/IPU), blijven relevant voor eenvoudige toepassingen of als educatief instrument, maar zijn minder robuust bij complexe randvoorwaarden. Simulatieve of probabilistische benaderingen hebben meerwaarde bij scenarioverkenningen of in geavanceerde agent-based modellen, maar zijn voor structurele toepassing binnen SIVMO vooralsnog minder geschikt.

Voor de Nederlandse situatie ligt een logische route in het selecteren en aanpassen van een open, optimalisatie-gebaseerde synthesemethode als generiek startpunt, gecombineerd met ruimte voor aanvullende modules of methoden in specialistische domeinen. De behoefte aan transparantie, validatie en onderhoudbaarheid pleit daarbij voor een open source aanpak, liefst met een brede gebruikersgemeenschap.

6
solutions

6 Richting en aanbevelingen

Dit hoofdstuk vertaalt de bevindingen uit de inventarisatie, interviews en beoordeling naar keuzes en randvoorwaarden voor de Nederlandse modelpraktijk. Het schetst welke functies een populatiesynthese moet kunnen vervullen, en welke eisen dat stelt aan data, privacy, reproduceerbaarheid en beheer. Op basis daarvan worden ontwikkelpaden en de organisatie besproken, inclusief samenwerking met leveranciers, inzet van open source, en een beheer- en governance-aanpak die voor gebruik en doorontwikkeling.

6.1 Uitgangspunten

Dit hoofdstuk bouwt voort op de analyse van methoden en tools en op de interviews met Nederlandse en internationale partijen. Het doel is niet om één methode of tool voor te schrijven, maar om een richting te schetsen voor een duurzame en breed gedragen toepassing van populatiesynthese binnen SIVMO. Daarbij zijn de volgende uitgangspunten leidend.

Ten eerste is er behoefte aan **brede en consistente inzetbaarheid**. Binnen SIVMO wordt gewerkt met diverse transportmodellen, waaronder trip-based, tour-based en in de toekomst ook activity-based modellen. Een landelijke populatiesynthese moet daarom niet op één specifiek modeltype zijn toegesneden, maar een generieke basis bieden die in verschillende modelcontexten kan worden toegepast zonder dat per toepassing een aparte populatie hoeft te worden gegenereerd. Dit vraagt om methoden die stabiel omgaan met meerdere schaalniveaus en uiteenlopende randvoorwaarden.

Ten tweede moet de oplossing **aansluiten op de bestaande Nederlandse modelpraktijk en data-infrastructuur**. Populatiesynthese staat niet op zichzelf, maar is ingebed in een omgeving met CBS-statistieken, ODiN, demografische prognoses en bestaande zoneringen. Uit de interviews blijkt dat de praktische uitvoerbaarheid sterk wordt bepaald door privacyregels, schaalbeperkingen en beschikbaarheid van marges. Een werkbare oplossing moet deze randvoorwaarden respecteren en hier methodologisch op zijn ingericht.

Ten derde is een **combinatie van reproduceerbaarheid en gecontroleerde flexibiliteit** noodzakelijk. Voor beleidsanalyses en prognoses is het een vereiste dat dezelfde invoer leidt tot dezelfde uitkomsten, zodat resultaten verifieerbaar en vergelijkbaar zijn. Tegelijkertijd moet de populatiesynthese voldoende flexibel zijn om te kunnen omgaan met scenario's, nieuwe databronnen of aangepaste indelingen. Dit betekent dat variatie mogelijk moet zijn, maar altijd expliciet, gedocumenteerd en beheerst.

Ten vierde is **transparantie, overdraagbaarheid en toekomstbestendigheid** een expliciet uitgangspunt. De interviews laten zien dat afhankelijkheid van gesloten systemen of persoonsgebonden kennis risico's met zich meebrengt voor continuïteit en kwaliteitsborging. Voor SIVMO is het daarom belangrijk dat de gekozen richting gebaseerd is op open specificaties, heldere documentatie en een beheerstructuur die overdraagbaar is tussen organisaties en in de tijd. Dit geldt niet alleen voor de software zelf, maar ook voor aannames, marges, validatiestappen en versiebeheer.

Deze uitgangspunten vormen samen het beoordelingskader voor de mogelijke routes die in de volgende paragraaf worden besproken. Zij maken duidelijk dat de keuze voor een populatiesynthese binnen SIVMO primair een **strategische en organisatorische afweging** is, en pas in tweede instantie een keuze tussen specifieke methoden of tools.

6.2 Mogelijke routes voor implementatie

Op basis van de analyse in de voorgaande hoofdstukken kunnen drie hoofdroutes worden onderscheiden voor de verdere inrichting van populatiesynthese binnen SIVMO. Deze routes zijn niet strikt exclusief, maar representeren verschillende strategische keuzes met elk eigen consequenties voor beheer, transparantie en toekomstbestendigheid.

Route A – Standaardiseren van bestaande, reeds gebruikte tools
Deze route gaat uit van het aanwijzen van een synthesizer als dé standaard binnen SIVMO, bijvoorbeeld Octavius of SigPopu. Beide tools zijn ingebed in de Nederlandse modelpraktijk en hebben hun waarde bewezen in operationele toepassingen. Het voordeel van deze route is dat zij voortbouwt op bestaande kennis, data-indelingen en workflows, waardoor de implementatiedrempel relatief laag is. Voor regionale en nationale modellen kan dit op korte termijn efficiënt zijn.

Tegelijkertijd laten de interviews zien dat deze route beperkingen kent. De tools zijn niet open source en sterk gekoppeld aan specifieke leveranciers[13]. Dit beperkt de transparantie van methodologische keuzes en bemoeilijkt onafhankelijke door-ontwikkeling of aanpassing aan nieuwe modeltypen, zoals activity-based modellen. Bovendien ontstaat een structurele afhankelijkheid van externe partijen voor licenties, onderhoud, updates en validatie, wat op langere termijn risico's oplevert voor kennisborging binnen SIVMO.

Route B – Overnemen en aanpassen van een bestaande open synthesizer
Een tweede route is het selecteren van een bestaande open source synthesizer, zoals PopulationSim (en in mindere mate PopGen) en deze aanpassen aan de Nederlandse context. Dit houdt in: aansluiting op CBS-marges, ODiN-indelingen, Nederlandse zoneringen en expliciete borging van privacy- en schaalbeperkingen. De interviews

[13] In dit onderzoek is niet vastgesteld of deze tools zodanig kunnen worden ingekocht dat SIVMO-partners (mede-)eigenaar worden of broncode- en onderhoudsrechten verkrijgen. Als dat wel mogelijk is, kan dit een deel van de genoemde afhankelijkheden verminderen, maar dit vergt expliciete afspraken over eigenaarschap, wijzigingsrechten, documentatie, releasebeheer en continuïteit.

met DfT en PTV laten zien dat deze aanpak internationaal gangbaar is en alleen goed kan functioneren als governance, documentatie en validatie expliciet zijn ingericht.

Het belangrijkste voordeel van deze route is dat zij transparantie en overdraagbaarheid combineert met methodologische robuustheid. De openheid van de code maakt onafhankelijke toetsing en gezamenlijke doorontwikkeling mogelijk. Tegelijk vraagt deze route een initiële investering in afstemming, validatie en documentatie. Ook moet expliciet worden besloten welke onderdelen worden gestandaardiseerd en waar ruimte blijft voor project specifieke keuzes. Zonder centrale regie bestaat het risico dat varianten ontstaan die de reproduceerbaarheid ondermijnen.

Route C – Ontwikkelen van een nieuwe, generieke synthesizer
De derde route is het ontwikkelen van een nieuwe, modulair opgebouwde synthesizer die is ontworpen voor de Nederlandse situatie. Hierbij kan worden gedacht aan een architectuur waarin fitting, allocatie en validatie gescheiden modules vormen, op basis van vooraf gekozen methoden (bijvoorbeeld IPU, optimalisatie en sampling). Deze route biedt maximale flexibiliteit en maakt het mogelijk om vanaf het begin rekening te houden met privacy, schaalhiërarchie en toekomstige uitbreidingen.

De keerzijde is dat deze route de grootste investering vraagt, zowel in ontwikkeling als in structureel beheer. De interviews, onder andere met Significance en CBS, onderstrepen dat dergelijke systemen alleen duurzaam zijn wanneer eigenaarschap, financiering en onderhoud goed zijn belegd. Zonder institutionele borging bestaat het risico dat een nieuw systeem technisch sterk is, maar organisatorisch kwetsbaar blijft.

Vergelijkende beschouwing
De drie routes verschillen vooral in mate van openheid, beheersinspanning en ontwikkelvrijheid. Route A is pragmatisch en snel inzetbaar, maar beperkt toekomstvast en met afhankelijkheden. Route C is conceptueel aantrekkelijk, maar organisatorisch zwaar met hogere kosten. Route B vormt een tussenpositie: zij combineert openheid en reproduceerbaarheid met beheersbaarheid, mits de randvoorwaarden voor governance en standaardisatie expliciet worden ingevuld.

6.3 Aanbevolen richting

Keuze voor Route B als voorkeursrichting
Op basis van de literatuurstudie, de interviews en de beoordeling in hoofdstuk 5 komt Route B – het overnemen en aanpassen van een bestaande open synthesizer naar voren als de meest evenwichtige en toekomstgerichte richting voor SIVMO.

Deze route sluit het beste aan bij de in paragraaf 6.1 geformuleerde uitgangspunten. Zij combineert reproduceerbaarheid en transparantie met voldoende flexibiliteit om verschillende modeltypen te ondersteunen. Tegelijkertijd vermijdt zij de structurele afhankelijkheid van gesloten systemen, zonder de ontwikkel- en beheerrisico's van een volledig nieuwe tool te introduceren. Route B vraagt daarbij niet alleen een keuze voor een open synthesizer, maar ook voor de vorm van de output die als landelijke standaard gaat gelden.

Standaardoutput: gewichten versus volledig geïntegeriseerde populatie
Daarbij is een keuze nodig voor de vorm van de synthese: (i) een populatie op basis van een steekproef met ophoging (met gewichten) of (ii) een volledig geïntegeriseerde populatie met individuele huishoudens en personen. Een steekproef met ophoging volstaat voor veel toepassingen en is vaak lichter in beheer en gebruik. Een volledig geïntegeriseerde populatie ondersteunt microscopische activity- en agent-based modellering en maakt analyse op dynamisch te definiëren groepen mogelijk, wat de uitlegbaarheid kan vergroten. Deze variant vraagt wel om borging van privacy-risico's en om afspraken over schaal en detail.

Route B kan beide outputvormen ondersteunen door een steekproef-gebaseerde seed via weging en (waar gewenst) integerisatie om te zetten naar een volledige synthetische populatie. Dit vergroot de flexibiliteit voor analyse en uitlegbaarheid, maar stelt eisen aan schaalkeuzes, privacy- en herleidbaarheid-risico's en beheer van micro-output.

Privacy- en data scenario's voor SIVMO
De Nederlandse data- en privacycontext bepaalt in hoge mate welke varianten van populatiesynthese binnen SIVMO realistisch en organiseerbaar zijn. Voor de besluitvorming is het daarom nodig de randvoorwaarden expliciet te vertalen naar drie ontwerpkeuzes: (1) waar de synthese draait (binnen een gecontroleerde omgeving of daarbuiten), (2) welke outputvorm als standaard wordt gekozen (gewichten/ steekproef versus een volledig geïntegeriseerde populatie), en (3) op welk schaalniveau constraints en output realistisch zijn gegeven privacy- en publicatieregels.

Op basis hiervan onderscheiden we drie uitvoeringsscenario's:
1. *Basis binnen gecontroleerde omgeving* (max detail, beperkte deelbaarheid)
   De synthese (en eventuele integerisatie) vindt volledig plaats binnen een gecontroleerde omgeving. Dit scenario biedt de meeste ruimte voor detail en validatie, maar export en deelbaarheid van micro-output zijn beperkt.
2. *Basis buiten gecontroleerde omgeving* (breed reproduceerbaar, beperkt detail)
   De synthese draait buiten de gecontroleerde omgeving op basis van geaggregeerde input en constraints op een realistischer (grover) schaalniveau. Dit scenario is het meest reproduceerbaar voor een brede gebruikersgroep, maar beperkt het detailniveau van zowel constraints als output.
3. *Hybride* (reproduceerbare basis buiten + verrijking binnen gecontroleerde omgeving)
   Er wordt een reproduceerbare basispopulatie buiten de gecontroleerde omgeving ontwikkeld als landelijke standaard. Voor toepassingen die meer detail, micro-validatie of verrijking vereisen, vindt een aanvullende stap plaats binnen de gecontroleerde omgeving. Dit combineert brede toepasbaarheid met een route naar detail waar dat nodig is, mits versiebeheer en consistentie-afspraken strak zijn ingericht.

Consequenties voor de voorkeursrichting en het 'eindbeeld'
Gezien de ambitie om één landelijke standaard te hebben die bruikbaar is in zowel trip- en tour-based modellen als in (toekomstige) activity- en agent-based toepassingen, is een volledig geïntegeriseerde populatie inhoudelijk aantrekkelijk als eindbeeld. De haalbaarheid en organisatievorm daarvan hangen echter samen met het gekozen uitvoeringsscenario. In een hybride opzet (uitvoeringsscenario 3) ligt het

voor de hand om een breed reproduceerbare basis te combineren met een gecontroleerde route naar integerisatie en detail voor micro-toepassingen. Daarmee kan één uniforme basispopulatie de consistentie tussen ketens en analyses ondersteunen, terwijl de privacy- en beheerrandvoorwaarden expliciet zijn geborgd.

Onderbouwing en modulaire opzet
De interviews laten zien dat open optimalisatie-gebaseerde methoden, met PopulationSim als meest concrete referentie, methodologisch voldoende volwassen zijn voor landelijke toepassingen. Zowel PTV als DfT benadrukken dat de stabiliteit van dergelijke systemen primair wordt bepaald door de kwaliteit en consistentie van marges en door een zorgvuldig ingerichte workflow, niet door het specifieke optimalisatie-algoritme. Dit sluit goed aan bij de Nederlandse context, waarin CBS-marges, ODiN en andere openbare bronnen centraal staan en waarin privacy- en schaalbeperkingen structureel zijn.

Daarbij is wel van belang dat Route B ruimte laat voor een modulaire inrichting. Een open synthesizer kan worden opgezet met een duidelijke kern voor basisjaren, gebaseerd op deterministische optimalisatie, terwijl aanvullende modules later kunnen worden toegevoegd. Denk hierbij aan probabilistische componenten voor gevoeligheidsanalyses. Deze gelaagde opzet voorkomt 'over-ambitie' in een vroeg stadium en maakt gefaseerde doorontwikkeling mogelijk.

Implicaties voor standaardisatie en governance
De keuze voor outputvorm en uitvoeringsscenario moet vervolgens worden vastgelegd in standaarden (variabelen, marges, schaalniveaus en validatiestappen), zodat de landelijke populatie reproduceerbaar is en varianten beheersbaar blijven.

Een keuze voor Route B impliceert wel dat SIVMO verantwoordelijkheid neemt voor standaardisatie en governance. Dit betekent dat vastgelegd moet worden welke variabelen, marges en schaalniveaus onderdeel zijn van de landelijke standaard, hoe updates worden doorgevoerd en hoe validatie wordt georganiseerd. De interviews met CBS en TNO maken duidelijk dat zonder dergelijke afspraken de voordelen van een open aanpak verloren gaan door fragmentatie en uiteenlopende interpretaties.

Samenvattend wordt aanbevolen om PopulationSim, of een vergelijkbare open optimalisatie-gebaseerde synthesizer, te selecteren als uitgangspunt en deze gericht aan te passen aan de Nederlandse data- en modelpraktijk. Deze keuze biedt een robuuste basis voor brede toepassing binnen SIVMO, terwijl zij voldoende ruimte laat voor toekomstige uitbreidingen en methodologische vernieuwing.

6.4 Aandachtspunten en vervolgstappen

De keuze voor een open, optimalisatie-gebaseerde synthesizer als uitgangspunt vraagt om een zorgvuldige uitwerking. De interviews maken duidelijk dat succes minder afhangt van de gekozen methode dan van de wijze waarop data, processen en verantwoordelijkheden zijn georganiseerd. Daarom is het essentieel om de volgende aandachtspunten expliciet te adresseren.

Een eerste stap is het uitwerken van een heldere functionele specificatie. Deze moet vastleggen welke variabelen onderdeel zijn van de landelijke populatiesynthese, op welke geografische niveaus wordt gewerkt en welke consistentie-eisen gelden tussen personen, huishoudens en eventueel andere entiteiten. Daarbij is het belangrijk om onderscheid te maken tussen een kernset van variabelen die voor alle toepassingen beschikbaar is en aanvullende kenmerken die optioneel of projectspecifiek kunnen worden toegevoegd.

Vervolgens of tegelijkertijd is een toets op datatoegankelijkheid noodzakelijk. Dit betreft niet alleen de beschikbaarheid van microdata en marges uit CBS-publicaties en ODiN, maar ook de consistentie van definities, tijdreeksen en schaalniveaus. In dit stadium moet expliciet worden bepaald welke stappen binnen een gecontroleerde omgeving plaatsvinden en welke output buiten die omgeving mag worden gebruikt. Dit wordt bij voorkeur expliciet gekoppeld aan één van de uitvoeringsscenario's uit paragraaf 6.3 , zodat data- en privacykeuzes direct vertaalbaar zijn naar workflow, exporteerbaarheid en beheer. Inhoudelijke en procedurele afstemming met CBS is hierbij onmisbaar.

Een derde aandachtspunt is validatie. Naast standaard margecontroles moet worden vastgelegd welke aanvullende consistentietoetsen worden uitgevoerd, bijvoorbeeld het vergelijken van huishoudtotalen met persoonsaantallen of het controleren van plausibiliteit van combinaties van kenmerken. Deze validatiestappen moeten reproduceerbaar zijn en onderdeel vormen van de standaard workflow, zodat verschillen tussen runs verklaarbaar blijven.

Daarnaast is een pilotfase essentieel. Door de gekozen synthesizer eerst toe te passen op een beperkt (regionaal) gebied kan ervaring worden opgedaan met data-voorbereiding, uitvoering, validatie en interpretatie van resultaten. Deze pilot dient niet alleen om technische kinderziekten te identificeren, maar ook om gezamenlijk begrip te ontwikkelen bij SIVMO-partners over de mogelijkheden en beperkingen van de aanpak.

Een aandachtspunt betreft beheer en governance. Er moet duidelijkheid komen over wie verantwoordelijk is voor het onderhoud van de code, de actualisatie van data, de documentatie en de validatie van nieuwe versies. Dit vraagt om expliciete afspraken over eigenaarschap, besluitvorming en financiering. Daarbij is ook relevant hoe toekomstvast het samenwerkingsverband SIVMO is als beheerdrager. Wanneer het partnerschap in samenstelling of rol kan wijzigen, is het verstandig om beheer en eigenaarschap te beleggen bij een stabiele drager (bijvoorbeeld een aangewezen beheerorganisatie of een consortium met meerjarige afspraken), met heldere exit- en overdrachtsafspraken. Zonder dergelijke afspraken bestaat het risico dat de synthesizer versnipperd raakt of veroudert, zoals in eerdere internationale voorbeelden is gebleken.

Tot slot is het raadzaam om modulariteit vanaf het begin te verankeren. Door de kern van de populatiesynthese bewust beperkt en stabiel te houden, ontstaat ruimte om in latere fases aanvullende modules te ontwikkelen, bijvoorbeeld voor probabilistische analyses, scenario-specifieke uitbreidingen of koppelingen met transitiemodellen. Deze gefaseerde aanpak sluit aan bij de praktijkervaringen uit de interviews en vergroot de kans op duurzame toepassing binnen SIVMO.

Bijlage 1 Referenties

Albiston, Emily, David R. Lovell, Hao Wu & S. Travis Waller. 2024. "*A Neural Network Approach for Population Synthesis*." Simulation: Transactions of the Society for Modeling and Similation International 2024, Vol 100(8) 823-847. https://doi.org/10.1177/00375497241233597. albiston-et-al-2024-a-neural-network-approach-for-population-synthesis.pdf

Antoni, Jean-Philippe, Gilles Vuidel & Olivier Klein. 2017. "*Generating a Located Synthetic Population of Individuals, Households, and Dwellings*." LISER Working Paper No. 2017-07, February 2017. https://dx.doi.org/10.2139/ssrn.2972615 . generating-a-located-synthetic-population-of-individuals.pdf.

Balać, Miloš & Sebastian Hörl. 2021. "*Synthetic Population for the State of California Based on Open Data: Examples of San Francisco Bay Area and San Diego County*." Paper presented at the 100th Annual Meeting of the Transportation Research Board (TRB), Washington D.C. (virtual). HAL archive: https://hal.science/hal-03208848. TRB21_Cali.pdf.

Balakrishna, Ramachandran, Srinivasan Sundaram & Jim Lam. 2020. "*An Enhanced nd Efficient Population Synthesis Approach to Support Advanced Travel Demand Models*." Paper presented at the TRB 2020 Annual Meeting, Washington D.C., and submitted to Transportation Research Record. caliper-population-synthesis-trb-2020.pdf.

Beemster, Fieke. 2016. "*Developing a Population Synthesis Method Based on Lifestyles Towards Mobility for Travel Demand Modelling.*" MSc thesis, Delft University of Technology. Developing a Population Synthesis Method based on Lifestyles towards Mobility for Travel Demand Modelling.pdf.

Borysov, Stanislav S., Jeppe Rich & Francisco Camara Pereira. 2019. "*How to Generate Micro-Agents? A Deep Generative Modeling Approach to Population Synthesis.*" Transportation Research Part C: Emerging Technologies. 106: 73–97. https://doi.org/10.1016/j.trc.2019.07.006 . scalable_population_synthesis_rev1_clean_1.pdf.

Borysov, Stanislav S., Jeppe Rich & Francisco C. Pereira. 2019. "*Scalable Population Synthesis with Deep Generative Modeling*." Transportation Research Part C: Emerging Technologies 106(2019), pp. 73-79. https://doi.org/10.1016/j.trc.2019.07.006. 180806910.pdf

Borysov, Stanislav S., Jeppe Rich & Francisco C. Pereira. n.d. "*Population Synthesis Meets Deep Generative Modelling*". Lyngby: Department of Management Engineering, Technical University of Denmark. 5380.pdf.

Chapuis, Kevin & Patrick Taillandier. 2019. *A Brief Review of Synthetic Population Generation Practices in Agent-Based Social Simulation*. Conference paper, September 2019. https://www.researchgate.net/publication/335601121 . A_review_of_synthetic_population_generation_process_in_social_simulation-4.pdf.

Choupani, Abdoul-Ahad & Amir Reza Mamdoohi. 2016. *"Population Synthesis Using Iterative Proportional Fitting (IPF): A Review and Future Research."* Transportation Research Procedia 17: 223–233. https://doi.org/10.1016/j.trpro.2016.11.078. 1-s20-S2352146516306925-

main.pdf. Population_Synthesis_Using_Iterative_Proportional_ (1).pdf & 1-s20-S2352146516306925-main.pdf

Feng, Lewen & Md. Kamruzzaman. 2023. "*Comparing Major Population Synthesis Techniques: A Case Study in Monash, Victoria*." Paper presented at the Australasian Transport Research Forum 2023, Perth, Australia, November 29 – December 1, 2023. ATRF_2023_Paper_114.pdf.

Fournier, Nicholas, Eleni Christofa, Arun Prakash Akkinepally & Carlos Lima Azevedo. 2018. "*An Integration of Population Synthesis Methods for Agent-Based Microsimulation*." Extended abstract submitted to the 97th Annual Meeting of the Transportation Research Board (TRB), Washington, D.C. CLA_Population_TRB_2018_extabs.pdf.

Hafezi, Mohammad Hesam & Muhammad Ahsanul Habib. 2014. *"Synthesizing Population for Microsimulation-Based Integrated Transport Models Using Atlantic Canada Micro-Data*." Procedia Computer Science 37: 410–415. https://doi.org/10.1016/j.procs.2014.08.061. 1-s20-S1877050914010266-main.pdf

Harland, Kirk, Alison Heppenstall, Dianna Smith & Mark Birkin. 2012. *Creating realistic synthetic populations at varying spatial scales: A comparative critique of population synthesis techniques*. Journal of Artificial Societies and Social Simulation, 15 (1). 1. ISSN 1460-7425 . https://doi.org/10.18564/jasss.1909 . Creating Realistic Synthetic Populations_published.pdf

Hörl, Sebastian & Miloš Balać. 2020. *"Open Data Travel Demand Synthesis for Agent-Based Transport Simulation: A Case Study of Paris and Île-de-France"*. Working Paper, Transport and Mobility Laboratory, EPFL. ab1499.pdf.

Jain, Shubham, Nicole A. Ronald & Stephan Winter. 2015. "*Creating a Synthetic Population: A Comparison of Tools.*" Paper presented at the 3rd Conference of Transportation Research Group of India (CTRG), December 2015. https://www.researchgate.net/publication/291608775. 587-CameraReady.pdf.

Joemanbaks, Shaya Q. J. & Jan Kiel. 2022. *The Potential Of Openstraatmap Data In Transport Models: A Case Study In Zoetermeer.* Paper presented at the European Transport Conference 7-9 September 2022, Milan. ETC The Potential of OpenStreetMap Data in Transport Models - A Case Study in Zoetermeer - Joemmanbaks & Kiel.pdf

Kagho, Grace O., Anugrah Ilahi, Miloš Balać & Kay W. Axhausen. 2020. *"Synthetic Population of Greater Jakarta: An Iterative Proportional Updating Approach."* Paper presented at the 20th Swiss Transport Research Conference (STRC), Ascona, Switzerland, May 13–15, 2020. Kagho_EtAl.pdf.

Kang, Jaewoong, Young Kim, Muhammad Mu'az Imran, Gi-sun Jung & Yun Bae Kim. 2023. "*Generating Population Synthesis Using a Diffusion Model.*" In Proceedings of the 2023 Winter Simulation Conference, edited by C.G.. Corlu et al., 2944–2955. IEEE. 2023-kang-kim-imran-jung-kim.pdf.

Kouwenhoven, Marco & Dylan Mulders. 2022. *Opzet populatiesimulator*. Memo 20025-M27 v6, June 7, 2022. The Hague: Significance. M27_-_opzet_populatiesimulator_-_v9.pdf.

Kukic, Marija & Michel Bierlaire. 2021. "*The Case of Population Synthesis at the Level of the Households*." Paper presented at the 21st Swiss Transport Research Conference (STRC), Ascona, September 12–14, 2021. Kukic_Bierlaire.pdf.

La, Duc Minh & Hai L. Vu. 2024. "*A Pool-Based Approach to Population Synthesis in Transport Modeling."* Paper to be presented at the Australasian Transport

Research Forum 2024, Melbourne, Australia, November 27–29, 2024. ATRF2024_Abridged_51-1.pdf.

Lim, Poh Ping & David Gargett. 2013. *"Population Synthesis for Travel Demand Forecasting."* In Proceedings of the 36th Australasian Transport Research Forum (ATRF), October 2–4, 2013, Brisbane, Australia. 2013_lim_gargett.pdf.

MARG. 2016. *PopGen: Synthetic Population Generator* [online]. Mobility Analytics Research Group. Available at http://www.mobilityanalytics.org/popgen.html, Accessed 20250716.

Müller, Kirill. 2014. *A Generalized Approach to Population Synthesis*. PhD diss., ETH Zürich. https://doi.org/10.3929/ethz-b-000171586. phd-thesis-muelleki-final.pdf.

Müller, K. & Kay W. Axhausen. 2010. "*Population Synthesis for Microsimulation: State of the Art*." Paper presented at the Swiss Transport Research Conference (STRC), ETH Zürich, August 2010. https://doi.org/10.3929/ethz-a-006127782. eth-1623-01.pdf & Mueller.pdf.

Nowok, Beate & Chris Dibben, 2018. "*Putting synthetic people in place: creating synthetic data for spatial analysis at the individual level.* QCumber-EnvHealth project: WP3 health data)

Nowok, Beata, Gillian M. Raab, Chris Dibben, 2016. *Synthpop: Bespoke Creation of Synthetic Data in R*. Journal of Statistical Software, 74(11), 1-26.

Paul, Binny Mathew, Jeff Doyle, Ben Stabler, Joel Freedman, Alex Bettinardi, 2018. "*Multi-level Population Synthesis Using Entropy Maximization-Based Simultaneous List Balancing"*. Transportation Research Board 97th Annual Meeting. Washington, 2018-1-7 – 2018-1-11.

Rahman, Md. Nobinur & Mahmudur Rahman Fatmi. 2023. "*Population Synthesis Accommodating Heterogeneity: A Bayesian Network and Generalized Raking Technique*." Transportation Research Record 2677 (6): 41–57. https://doi.org/10.1177/03611981221144289 rahman-fatmi-2023-population-synthesis-accommodating-heterogeneity-a-bayesian-network-and-generalized-raking-technique.pdf

Rich, Jeppe. 2018. "*Large-Scale Spatial Population Synthesis for Denmark.*" European Transport Research Review 10 (63). https://doi.org/10.1186/s12544-018-0336-2 . s12544-018-0336-2.pdf.

Rich, Jeppe, Gunnar Flötteröd, Sergio Garrido & Francisco Pereira. 2019. *Review of Population Synthesis Methodologies*. Paper presented at the hEART 2019 conference. Department of Management Engineering, Technical University of Denmark. hEART_2019_paper_122.pdf

RSG, 2021. "MWCOG Population Synthesizer. Final Report". Prepared for the Metropolitan Washington Council of Governments (MWCOG). https://www.mwcog.org/assets/1/6/MWCOG_Population_Synthesizer_COG_final1.pdf

Saadi, Ismaïl, Hamed Eftekhar, Jacques Teller & Mario Cools. 2016. "*Investigating the Scalability in Population Synthesis: A Comparative Approach*." Transportation Planning and Technology 39(6): 569–591. https://orbi.uliege.be/bitstream/2268/229325/1/_system_appendPDF_proof_hi%20%281%29.pdf. _system_appendPDF_proof_hi (1).pdf

Significance. 2020. *Toetsingskader Nieuw Nationaal Personenautoparkmodel*. Memo aan Konstanze Winter, Remko Smit en Jordy van Meerkerk, 10 november 2020. Den Haag: Significance. M02 - Toetsingskader v04.pdf.

Significance. 2021. *Backcast LMS: Vergelijking prognose en waargenomen ontwikkeling.* Rapportnummer 21027. Den Haag: Significance.

Significance. 2022a. QUAD en GWI: Resultaten Fase I. Rapportnummer 22050. Den Haag: Significance. 22050 R01 QUAD en GWI - Fase I versie 4.pdf.

Significance. 2022b. *QUAD en GWI: Fase Resultaten Fase II*. Rapportnummer 22050. Den Haag: Significance. 22050 R02 QUAD en GWI - Fase II.pdf

Significance. 2024a. *Vergelijking QUAD – SigPopu*. Intern rapport, versie februari 2024. Den Haag: Significance. Vergelijking QUAD - SigPopu - v5.pdf

Significance. 2024b. *Actualisatie invoer huishoudsimulator*. Memo 24010-M02 v8, 22 mei 2024. Den Haag: Significance. 24010-M02 - Actualisatie invoer huishoudsimulator v08.pdf

Wu, Hao & Cheng Lyu. 2024. *Simulation-Based Comparative Analysis of Synthetic Population Generation Methods: A Framework for Travel Diary Validation in Transport Simulation*. Chair of Transportation Systems Engineering, TUM School of Engineering and Design, Technische Universität München.

Wu, Hao & Cheng Lyu. 2024. *Tabular Data Imputation for Synthetic Population with Diffusion Models*. Chair of Transportation Systems Engineering, TUM School of Engineering and Design, Technische Universität München.

Ye, Xin, Karthik C. Konduri, Ram M. Pendyala, Bhargava Sana & Paul Waddell. 2009. *"A Methodology to Match Distributions of Both Household and Person Attributes in the Generation of Synthetic Populations."* Paper gepresenteerd op de 88e Annual Meeting of the Transportation Research Board, Washington, D.C., januari 2009.

Bijlage 2 Literatuur review

2.1 Inleiding

Voor dit onderzoek is een systematische literatuurreview uitgevoerd naar bestaande methoden en toepassingen van population synthesis (populatiesynthese), met de nadruk op de inzet binnen transport- en mobiliteitsmodellen. De literatuur is in eerste instantie verzameld via online zoekopdrachten op de termen *population synthesis transport model*. Hierbij is gebruik gemaakt van vrij toegankelijke bronnen, waaronder zoekmachines, het platform ResearchGate en het AI-gebaseerde zoekinstrument Elicit. Deze aanpak leverde een brede reeks documenten op, waaronder wetenschappelijke artikelen, conferentiepapers, werkdocumenten, interne memo's en scripties.
Aanvullend is de literatuur verrijkt met informatie die door Rijkswaterstaat is aangeleverd over twee modellen die in Nederland worden toegepast in verkeers- en vervoermodellen: QUAD en SigPopu. Beide modellen zijn inhoudelijk beoordeeld op basis van de aangeleverde rapportages.
In totaal zijn 34 documenten geselecteerd en geanalyseerd. Elk document is geëvalueerd op volledigheid, bruikbaarheid en relevantie voor de Nederlandse context. Waar nodig zijn publicatiegegevens (zoals titel, auteurs, jaartal en type publicatie) aangevuld of gecorrigeerd op basis van de originele documenten (PDF en/of tekstbestand).

2.2 Analysekader en gehanteerde criteria

De literatuur is gestructureerd geanalyseerd aan de hand van een vast format, waarin voor elk stuk drie onderdelen zijn opgenomen:

- **Titelgegevens**
  Volledige titel, auteurs, jaar van publicatie, publicatietype (artikel, rapport, paper, scriptie).
- **Samenvatting**
  Een korte, feitelijke beschrijving van het doel van de publicatie, de gebruikte methode(n), de context waarin het onderzoek is uitgevoerd, en de belangrijkste resultaten of inzichten.
- **Analyse op acht inhoudelijke criteria**
  Elke publicatie is beoordeeld op:
  - **Toegepaste methode(n)**. Welke techniek(en) worden gebruikt voor synthese?
  - **Gebruikte inputdata**. Welke databronnen vormen de basis?
  - **Kenmerken van de populatie**. Gaat het alleen om personen en huishoudens of ook om voertuigen, activiteiten, locaties?
  - **Schaalniveau**. Nationaal, regionaal of lokaal?
  - **Output**. Wat levert de methode concreet op (bv. microdata, gedisaggregeerde populaties)?

- **Validatie**. Hoe is de kwaliteit van de synthetische populatie gecontroleerd?
- **Openbaarheid**. Is de data, methode of tool vrij beschikbaar of gesloten?
- **Toepasbaarheid voor SIVMO**. Wat is de relevantie van de bron voor SIVMO?

De individuele analyses zijn opgenomen in de bijlage. De uitkomsten vormen de basis voor een vergelijkend overzicht van benaderingen, waarmee een onderbouwde beoordeling mogelijk is van de geschiktheid van bestaande methoden en systemen voor populatiesynthese in de Nederlandse modelpraktijk.

Albiston, Emily, David R. Lovell, Hao Wu & S. Travis Waller. 2024.
"*A Neural Network Approach for Population Synthesis.*" Simulation: Transactions of the Society for Modeling and Similation International 2024, Vol 100(8) 823-847. https://doi.org/10.1177/00375497241233597. albiston-et-al-2024-a-neural-network-approach-for-population-synthesis.pdf

Samenvatting

Het artikel onderzoekt hoe verschillende technieken voor populatiesynthese presteren bij het genereren van synthetische populaties voor verkeerstoepassingen. Omdat microsimulatiemodellen gedetailleerde informatie op individueel niveau vereisen, maar zulke data niet direct beschikbaar zijn, worden synthetische populaties gegenereerd op basis van geaggregeerde statistieken. Albiston et al. vergelijken in dit kader zes methoden: baseline sampling, direct sampling, iteratieve proportionele fitting (IPF, in twee varianten), Bayesian networks, en een artificial neural network (ANN)-benadering. Daarbij worden ook meerdere evaluatiematen toegepast, waaronder traditionele foutmaten (RMSE, MAE) en set-theoretische maten (Jaccard-index, intersection rate).

De studie toont aan dat geen enkele foutmaat op zichzelf toereikend is: klassieke maten zoals RMSE geven vertekende resultaten bij hoge dimensionaliteit en kleine populaties. De auteurs introduceren daarom aanvullende vergelijkingsmaten op basis van archetype-dekking. Daarmee onderscheiden ze fouten die ontstaan door een verkeerde verdeling van bestaande archetypes van fouten die ontstaan door het genereren van niet-bestaande archetypes. Deze benadering levert beter inzicht in de kwaliteit van de gegenereerde populaties.

Uit de experimentele resultaten blijkt dat IPF het best presteert in scenario's met relatief grote populaties en weinig variabelen. In situaties met meer kenmerken of kleinere populaties neemt de nauwkeurigheid van IPF echter sterk af. Het ANN-model presteert dan relatief beter, vooral in termen van het vermijden van foutieve archetypes. Direct sampling toont zich robuust en consistent, met goede resultaten in verschillende omstandigheden en zeer korte rekentijd.

De auteurs concluderen dat ANN-technieken beloftevol zijn, maar veel rekentijd vergen. IPF blijft geschikt bij rijke data, maar kent beperkingen bij populatieschaarste. Nieuwe evaluatiemethoden bieden een krachtiger analyse-instrument en dragen bij aan beter geïnformeerde keuzes in populatiesynthese voor verkeerstoepassingen.

Analyse
1. Methode. Het artikel vergelijkt de volgende populatiesynthesetechnieken: baseline sampling, direct sampling, IPF, Bayesian networks en een artificial neural network (ANN). De aanpak is systematisch en experimenteel van aard. De auteurs bouwen synthetische populaties op met verschillende combinaties van kenmerken, variëren populatiegrootte (van dense tot zeer sparse) en toetsen prestaties onder uiteenlopende condities. De methode is uitgebreid beschreven, inclusief specificaties van variabelecombinaties en zone-indelingen.

2. Inputdata. De gebruikte data zijn afkomstig uit de UK Census 2011 (Local Authority Microdata), die een 5%-steekproef bevat van elk van de 265 regio's. De auteurs

simuleren kleinere zones binnen deze regio's. Zowel gecontroleerde als ongecontroleerde variabelen worden geselecteerd uit 121 persoonskenmerken. De inputdata zijn dus openbaar, representatief en realistisch voor transporttoepassingen.

3. Kenmerken. Het model genereert microdata op persoonsniveau, waarbij archetypes worden opgebouwd op basis van categorische kenmerken. Het model kan omgaan met variërende aantallen variabelen, wat schaal en complexiteit beïnvloedt. De studie richt zich op synthetische populatiebouw, niet op gedrag of verplaatsingen. Er is geen expliciete koppeling naar huishoudens, hoewel dat in vervolgonderzoek wordt gesuggereerd.

4. Schaal. De schaal is aanpasbaar: het model wordt getest op zones met uiteenlopende bevolkingsaantallen (van 169 tot ruim 10.000). Dit maakt de vergelijking tussen technieken bij verschillende niveaus van populatieschaarste mogelijk. De schaal is dus zowel lokaal (wijkniveau) als regionaal toepasbaar, al is het model getraind op referentieregio's binnen het VK.

5. Output. De output bestaat uit synthetische populaties (personen met kenmerkcombinaties). De kwaliteit wordt gemeten met error rates en met set-based vergelijkingen (zoals Jaccard-similariteit). Er wordt niet doorgerekend naar verkeersmodellen, dus geen directe gedrags- of mobiliteitsuitkomsten. De synthese dient als input voor activity-based modellen.

6. Validatie. Validatie gebeurt op basis van meerdere maten: RMSE, MAE, population error rate, intersection rate, en difference rate. De auteurs laten zien dat traditionele foutmaten tekortschieten bij hoge dimensionaliteit. Ze onderbouwen het gebruik van set-based vergelijkingen. Validatie vindt plaats door vergelijking met de 'ground truth' (de oorspronkelijke microdata). Er is geen externe of gedragsmatige validatie.

7. Openbaarheid. De onderliggende data zijn publiek beschikbaar (UK Census 2011), en de methoden zijn beschreven met gebruik van open R-pakketten. De code is echter niet meegestuurd of publiek beschikbaar gesteld. Het model is daardoor *deels* reproduceerbaar, mits voldoende programmeerkennis. De ANN-structuur is volledig beschreven, maar trainingsbestanden ontbreken.

8. Toepasbaarheid voor SIVMO. Het artikel is relevant voor SIVMO vanwege de grondige vergelijking van methoden en de duidelijke beperkingen van IPF bij kleine populaties of veel kenmerken. ANN wordt gepresenteerd als kansrijke techniek, met het nadeel van lange rekentijd. De koppeling met mobiliteitsmodellen ontbreekt echter. Aanpassing voor huishoudniveau, multimodaal gebruik of gedragsmodellen is nodig. De aanpak biedt waardevolle lessen voor de keuze van synthetische technieken in Nederland, mits verder aangepast aan de Nederlandse context.

Antoni, Jean-Philippe, Gilles Vuidel & Olivier Klein. 2017.
"*Generating a Located Synthetic Population of Individuals, Households, and Dwellings."* LISER Working Paper No. 2017-07, February 2017.
https://dx.doi.org/10.2139/ssrn.2972615 . generating-a-located-synthetic-population-of-individuals.pdf.

Samenvatting

Dit artikel beschrijft de ontwikkeling van de MobiSim Population Synthesizer: een model en softwarepakket om een gesynthetiseerde, ruimtelijk gelokaliseerde populatie te genereren op individueel, huishoudelijk en woningniveau. Deze synthetische populaties zijn essentieel voor moderne stadsmodellen zoals activity-based models (ABM) en multi-agent systems (MAS), die behoefte hebben aan gedetailleerde inputdata op microniveau. Omdat zulke microdata meestal niet beschikbaar zijn vanwege privacybeperkingen, biedt deze methode een manier om uit geaggregeerde sociaaleconomische en geografische data een plausibele, maar anonieme populatie te genereren.

De methode bestaat uit twee stappen: eerst wordt een gesynthetiseerde bevolking gegenereerd uit geaggregeerde data, vervolgens wordt deze populatie ruimtelijk gelokaliseerd op basis van gedetailleerde geografische databronnen zoals BD-Topo. Dit gebeurt op het niveau van gebouwen en zelfs op verdiepingshoogte. De procedure houdt rekening met kenmerken zoals leeftijd, geslacht, huishoudstructuur en woninggrootte. Via probabilistische toewijzing en optimalisatietechnieken (zoals simulated annealing) worden individuen gegroepeerd in huishoudens en toegewezen aan woningen en gebouwen.

De methode is getest op drie Franse steden (Besançon, Lille en Strasbourg) en levert realistische resultaten op. Validatie vindt plaats door vergelijking met officiële censusdata. Hoewel enkele afwijkingen optreden voor kleinere populatiecategorieën (zoals eenpersoonshuishoudens of ouderen), blijkt de variatie in het model beperkt en de overall foutmarges klein.

Tot slot biedt het model een bruikbare basis voor microsimulaties en gedragssimulaties in een stedelijke context. Hoewel de huidige versie een statische representatie oplevert, opent de structuur perspectieven voor dynamische uitbreidingen. De methode is toepasbaar op alle gebieden met vergelijkbare census- en geografische databronnen, wat het potentieel vergroot voor internationale toepassing in ruimtelijk-georiënteerde simulatiemodellen.

Analyse

1. Methode. De populatie wordt gesynthetiseerd in twee stappen. Eerst worden agenten, huishoudens en woningen gegenereerd op basis van geaggregeerde sociaaleconomische censusdata. Vervolgens worden deze elementen ruimtelijk gepositioneerd in gebouwen met behulp van gedetailleerde geografische data (BDTopo). De toewijzing gebeurt probabilistisch, waarbij onder andere simulated annealing wordt toegepast voor het optimaliseren van gezinssamenstelling en woningtoewijzing. De methode is expliciet ontwikkeld om te functioneren zonder microdata en is toepasbaar in landen waar alleen geaggregeerde data beschikbaar zijn.

2. Inputdata. Het model gebruikt twee hoofdbronnen:
Geaggregeerde sociaaleconomische data van het Franse nationale bureau voor statistiek (INSEE), beschikbaar op het IRIS-niveau (~16.000 zones).
Gedetailleerde vector-geodata van het Franse kadaster (BDTopo), met per gebouw hoogte, volume en locatie. De gebruikte data zijn publiek toegankelijk.

3. Kenmerken. De gegenereerde agenten hebben kenmerken als leeftijd, geslacht, opleidingsniveau en arbeidsstatus. Huishoudens worden ingedeeld naar type (bijv. alleenstaand, koppel met kinderen), en woningen hebben attributen zoals aantal kamers, type (appartement of huis), en ruimtelijke locatie (inclusief verdiepingshoogte).

4. *Schaal.* Het model werkt op het niveau van Franse steden, met drie casussen: Besançon, Lille en Strasbourg. De populatie wordt gekoppeld aan individuele gebouwen en verdiepingen, waardoor een fijnmazige ruimtelijke resolutie ontstaat. Invoergegevens zijn beschikbaar voor heel Frankrijk, waardoor nationale toepassing mogelijk is.

5. Output. De output bestaat uit drie gekoppelde databestanden: agenten, huishoudens en woningen, inclusief hun locatie. Deze kunnen gekoppeld worden aan GIS-data voor simulaties en analyses. Resultaten zijn geschikt voor invoer in agent-based of microsimulatiemodellen en zijn geverifieerd op aggregatieniveau.

6. Validatie. Het model is gevalideerd met twee tests: een stabiliteitstest (100 runs met gelijke parameters) en een validatietest (vergelijking van synthetische output met censusdata op IRIS-niveau). Fouten zijn beperkt (<1% op de meeste variabelen), met grotere afwijkingen bij kleine categorieën (bv. ouderen, grote gezinnen). Validatie op individueel niveau is niet mogelijk vanwege het ontbreken van referentiegegevens.

7. Openbaarheid. Het model (MobiSim Population Synthesizer) is vrij beschikbaar als downloadbare Java-software (https://sourceforge.net/projects/popsynthe/). Ook de gebruikte datasets (INSEE en BDTopo) zijn publiek toegankelijk in Frankrijk. Documentatie en voorbeelddata zijn via de projectwebsite beschikbaar.

8. Toepasbaarheid voor SIVMO. De methode is goed bruikbaar in contexten zoals SIVMO, mits geaggregeerde bevolkingsdata en gedetailleerde geodata beschikbaar zijn. De nadruk op locatie en koppeling aan gebouwen sluit goed aan bij toepassingen in ruimtelijke simulatie. Een beperking is de afwezigheid van dynamiek (verhuizingen, demografie), waardoor het model statisch blijft. Bij uitbreiding met dynamische componenten zou het model breder inzetbaar zijn.

Balać, Miloš & Sebastian Hörl. 2021.
"*Synthetic Population for the State of California Based on Open Data: Examples of San Francisco Bay Area and San Diego County.*" Paper presented at the 100th Annual Meeting of the Transportation Research Board (TRB), Washington D.C. (virtual). HAL archive: https://hal.science/hal-03208848. TRB21_Cali.pdf.

Samenvatting
Dit paper presenteert een reproduceerbare en open-source aanpak voor het genereren van synthetische populaties, met toepassing op twee regio's in Californië: de San Francisco Bay Area en San Diego County. De methode is ontworpen als invoer voor agent-based verkeersmodellen zoals MATSim, en bestaat volledig uit vrij toegankelijke componenten. Daarmee richt het project zich expliciet op transparantie en overdraagbaarheid.

De werkwijze omvat meerdere stappen. Allereerst wordt een synthetische populatie gegenereerd met behulp van PopGen, een IPF-gebaseerde tool die huishoudens en personen opbouwt op basis van Amerikaanse censusdata en de American Community Survey (ACS). De gegenereerde populatie omvat basiskenmerken zoals leeftijd, geslacht, inkomen, werkstatus en autobezit. Vervolgens worden deze individuen verrijkt met dagelijkse activiteitenschema's (activity chains), door middel van hot-deck statistical matching met data uit de California Household Travel Survey (CHTS). Hierbij wordt gematcht op onder andere leeftijd, geslacht, werkstatus en bereikbaarheidskenmerken.

Daarna volgen meerdere imputatiestappen: eerst wordt een thuislocatie toegekend, vervolgens worden werk- en onderwijslocaties bepaald op basis van plausibele verplaatsingspatronen, en tot slot worden locaties voor niet-verplichte activiteiten geselecteerd op basis van afstandsdistributies. Deze stappen gebruiken onder andere data uit OpenStreetMap en het ministerie van Onderwijs.

De validatie laat zien dat de populatie op hoofdlijnen goed aansluit bij bekende verdelingen. In de Bay Area zijn de sociaaleconomische verdelingen nagenoeg identiek aan de inputmarges. In San Diego County zijn er enkele afwijkingen, vooral bij activiteitentypes, wat deels te verklaren is door verschillen tussen de CHTS en de ACS. De auteurs erkennen deze beperkingen, maar stellen dat de gegenereerde populaties bruikbaar zijn voor scenarioverkenningen en beleidsanalyse.

Analyse

1. Methode. De methode is gebaseerd op een open-source pipeline voor synthetische populatiegeneratie, met als kern het PopGen-tool (gebaseerd op iterative proportional fitting) en hot-deck matching voor activiteitentoewijzing. Locatie-imputatie gebeurt op basis van herkomst-bestemmingsmatrices (voor werk) en afstandsverdelingen (voor onderwijs en overige activiteiten), waarbij gebruik wordt gemaakt van heuristieken en KD-trees. De aanpak volgt een agent-based denkwijze waarbij huishoudens en individuen worden gegenereerd met sociaaleconomische en mobiliteitskenmerken, gevolgd door koppeling van dagelijkse activiteiten en locaties.

2. Inputdata. Alle gebruikte data zijn publiek beschikbaar, waaronder censusbestanden, ACS-microdata, CHTS-travelsurveys, OpenStreetMap en onderwijsinstellingsdata.

3. Kenmerken. De gegenereerde populatie bevat demografische kenmerken, huishoudsamenstelling, vervoermiddelenbezit, en volledige activiteitenschema's inclusief locaties en verplaatsingen.

4. Schaal. De toepassing richt zich op twee regio's in Californië (San Francisco en San Diego), met hoge resolutie (census tract-niveau). Het proces is schaalbaar naar andere gebieden in de VS, mits data beschikbaar zijn.

5. Output. De output bestaat uit een volledige microdata-populatie die direct inzetbaar is in MATSim-simulaties, inclusief dagelijkse routines, locaties en vervoermiddelenkeuze.

6. Validatie. De populatie is gevalideerd tegen censusmarges en activiteitendistributies. In de Bay Area is de aansluiting goed; in San Diego County zijn enkele afwijkingen zichtbaar, vooral bij school- en werkverplaatsingen door verschillen in survey jaren.

7. Openbaarheid. De gehele workflow is open-source en reproduceerbaar via eqasim pipeline. Alle code en documentatie zijn publiek beschikbaar (via GitHub).

8. Toepasbaarheid voor SIVMO. De aanpak is zeer bruikbaar voor SIVMO. Het combineert klassieke IPF met open-source uitbreidingen en gedragsimputatie, en is inzetbaar voor regio's waar open survey- en locatiedata beschikbaar zijn. De methode is modulair, transparant en direct toepasbaar in MATSim-gebaseerde beleidsmodellen. Mogelijke belemmering is het ontbreken van huishoudinteractie en beperkte detaillering van ruimtelijke context (aantrekkelijkheid van locaties), maar dit is uitbreidbaar binnen de open pipeline.

Balakrishna, Ramachandran, Srinivasan Sundaram & Jim Lam. 2020.
"*An Enhanced and Efficient Population Synthesis Approach to Support Advanced Travel Demand Models*." Paper presented at the TRB 2020 Annual Meeting, Washington D.C., and submitted to Transportation Research Record. caliper-population-synthesis-trb-2020.pdf.

Samenvatting

Het artikel introduceert een verbeterde methode voor het synthetiseren van populaties ter ondersteuning van geavanceerde verkeersvraagmodellen, met een focus op activity-based en hybride modellen. De auteurs signaleren dat conventionele methoden, vooral Iterative Proportional Fitting (IPF), weliswaar goed presteren op huishoudniveau, maar geen expliciete controle bieden over persoonskenmerken. Dit tekort belemmert de betrouwbaarheid van gedragsmodellen waarin individuele kenmerken zoals leeftijd en geslacht belangrijk zijn.

Om dit probleem te ondervangen, stellen de auteurs een aanpassing voor op de bestaande Iterative Proportional Updating (IPU) methode. Hun versie vermijdt het gebruik van volledig uitgeschreven joint distributions van huishoud- en persoonsvariabelen, wat het risico op 'zero-cells' en rekentijdproblemen aanzienlijk verkleint. In plaats daarvan worden afzonderlijke marges van variabelen gebruikt. Dit leidt tot kortere rekentijden en robuustere convergentie zonder aan nauwkeurigheid in te boeten.

De methode is geïmplementeerd in TransCAD en toegepast in twee regio's: Las Vegas (Nevada) en de Central Coast (Californië). In beide gevallen worden betere resultaten geboekt dan bij standaard-IPF: huishoudmarges worden even goed benaderd, maar persoonsmarges (leeftijd, geslacht) sluiten veel beter aan bij de doelverdelingen. Daarnaast is de rekenprestatie sterk verbeterd: beide casussen zijn in minder dan 10 minuten doorgerekend, waar eerdere methoden 11 tot 16 uur vergen.

Een belangrijk neveneffect dat de auteurs bespreken is de inconsistentie tussen huishoud- en persoonsmarges. Door externe databronnen te gebruiken (bijvoorbeeld retaildata) konden zij onrealistische marges corrigeren, in het bijzonder bij huishoudens met zeven of meer leden, waar censusdata onnatuurlijk hoge gemiddelden toonden. Deze correcties bleken essentieel voor een realistische populatie.

De auteurs concluderen dat hun aangepaste IPU-methode niet alleen accurater is, maar ook schaalbaar en geschikt voor praktijktoepassing binnen geavanceerde vervoersmodellen.

Analyse

1. Methode. De methode is gebaseerd op een verbeterde vorm van Iterative Proportional Updating (IPU), waarin zowel huishoud- als persoonskenmerken worden gematcht aan control totals (marginals). In tegenstelling tot de klassieke IPU wordt geen gezamenlijke kruistabel gebruikt, maar worden marges per variabele afzonderlijk gematcht. Dit vermindert het risico op lege cellen (zero-cell problem) en verkort de

rekentijd aanzienlijk. De methode is geïmplementeerd in TransCAD en getest in twee regio's.

2. Inputdata. Amerikaanse censusdata (PUMS) op PUMA-, block group- en blokniveau. Zowel huishoud- als persoonsmarges, waaronder inkomen, huishoudgrootte, voertuigen, leeftijd en geslacht.

3. Kenmerken. Zowel huishoud- als persoonskenmerken worden gesynthetiseerd. Voor huishoudens: grootte, inkomen, autobezit. Voor personen: leeftijd, geslacht, werksector. Uniek is dat persoonskenmerken expliciet en afzonderlijk worden opgenomen in het optimalisatieproces, in plaats van indirect via huishoudens.

4. Schaal. De methode ondersteunt geneste geografische schaalniveaus. IPU wordt toegepast op blokgroepniveau, IPF op blokniveau. Twee grootschalige Amerikaanse casestudy's: Las Vegas (13 PUMAs, 1294 blokgroepen, 24.521 blokken) en Central Coast (10 PUMAs, 941 blokgroepen, 39.660 blokken).

5. Output. De output is een synthetische populatie met huishoud- en persoonskenmerken die overeenkomen met de marges. De data kunnen direct worden gebruikt als input voor (activity-based) vervoermodellen. De methode is geïntegreerd in TransCAD, wat export in gangbare modelstructuren mogelijk maakt.

6. Validatie. Vergelijking tussen gegenereerde populaties en observaties op block group-niveau; grafische en numerieke analyse van marginale fits (voor en na toepassing van IPU). Daarnaast wordt gekeken naar rekenprestaties en consistentie van marges.

7. Openbaarheid. De methode is geïmplementeerd in TransCAD, een commercieel pakket. De code is niet open source, maar documentatie is publiek beschikbaar.

8. Toepasbaarheid voor SIVMO. De methode is bruikbaar voor SIVMO, mits toegang tot een pakket als TransCAD (een minpunt) of herimplementatie in open software. Sterk geschikt voor situaties waarin persoonskenmerken centraal staan in gedragsmodellering en waarin marges beschikbaar zijn op meerdere schaalniveaus.

Beemster, Fieke. 2016.
"*Developing a Population Synthesis Method Based on Lifestyles Towards Mobility for Travel Demand Modelling.*" MSc thesis, Delft University of Technology. Developing a Population Synthesis Method based on Lifestyles towards Mobility for Travel Demand Modelling.pdf.

Samenvatting
Deze masterthesis onderzoekt hoe de populatiesynthese in het mobiliteitsmodel Fountain kan worden verbeterd door het meenemen van zogenoemde 'lifestyles towards mobility' in plaats van traditionele socio-demografische kenmerken. Fountain gebruikt lifestyle-categorieën om reisgedrag te voorspellen, maar tot nu toe werd één uniforme verdeling van deze lifestyles voor heel Nederland aangenomen. De studie stelt dat dit tot fouten leidt, omdat de werkelijke verdeling tussen gebieden verschilt. Om dit aan te pakken, is een nieuwe populatiesynthesemethode ontwikkeld op basis van het Iterative Proportional Fitting (IPF) algoritme en een koppeling tussen leeftijd, geslacht en lifestylecategorieën.

De thesis bevat een literatuuronderzoek naar populatiesynthese-algoritmes, waarbij IPF als basis is gekozen. Daarnaast is een meta-analyse uitgevoerd op studies naar lifestyle en mobiliteit. Op basis van zeven relevante onderzoeken zijn vier overkoepelende lifestylegroepen geïdentificeerd, waarbij leeftijd en geslacht als belangrijkste voorspellers zijn geselecteerd. De oorspronkelijke studie van Anable (2011) is gebruikt om de koppeling tussen socio-demografische kenmerken en lifestyle in het model te leggen. Deze gegevens vormen de input voor het synthetiseren van een realistische populatie op lokaal niveau.

De nieuwe populatiesynthesetool bestaat uit drie stappen: het synthetiseren van de populatie met IPF Multizone, het toewijzen van lifestyles aan individuen op basis van hun leeftijd en geslacht, en het toewijzen van werkplekken via een bestaande OD-matrix. Twee case studies, in Amsterdam en Utrecht, tonen aan dat de verdeling van lifestyles tussen zones verschilt en dat het model met deze methode beter presteert, vooral bij het voorspellen van modaliteit en het vermijden van spitsuren.

De conclusie is dat het nieuwe model accurater is dan de oorspronkelijke Fountain-aanpak. Het resultaat is een robuustere, geografisch gedifferentieerde synthetische populatie die beter aansluit op gedragsverschillen en realistischer input levert voor mobiliteitsmodellen. Deze methode biedt daarmee een verbeterde basis voor beleidsanalyses en scenarioverkenningen.

Analyse

1. Methode. De studie past een aangepaste versie van Iterative Proportional Fitting (IPF) toe om een synthetische populatie te genereren, gebaseerd op leeftijd en geslacht. Deze populatie wordt vervolgens verrijkt met leefstijlkenmerken (lifestyles towards mobility), die via een meta-analyse zijn afgeleid uit zeven studies, waaronder Anable (2011). De methode kent drie stappen: IPF Multizone, leefstijltoekenning, en OD-matrixverdeling. Het eindresultaat is een set van zeven OD-matrices waarin elke persoon gekoppeld is aan een leefstijl.

2. Inputdata. De input bestaat uit socio-demoografische kenmerken (leeftijd, geslacht), marginaal beschikbare gegevens op zoneniveau voor Nederland, en surveydata over leefstijlen uit Anable (2011). Voor de OD-verdeling zijn bestaande verkeersmodellen gebruikt als referentie. Daarnaast is gebruikgemaakt van case-specifieke data voor Amsterdam (VMA) en Utrecht (spitsmijdenproject).

3. Populatiekenmerken. De populatie wordt gesynthetiseerd op leeftijd, geslacht en leefstijl. In de tool wordt expliciet rekening gehouden met ruimtelijke spreiding: verschillende zones krijgen verschillende leefstijlverdelingen toegewezen, gebaseerd op leeftijd- en geslachtsprofielen. Andere kenmerken (opleiding, inkomen, autobezit) zijn onderzocht maar uiteindelijk niet meegenomen wegens databeperkingen.

4. Schaal. De toepassing richt zich op het nationale schaalniveau (Nederland), met cases op stedelijk niveau (Amsterdam en Utrecht). De geografische eenheid betreft zones binnen steden, zoals stadsdelen en wijken. De temporele focus is statisch en richt zich niet op een specifieke prognosehorizon.

5. Output. De output bestaat uit een synthetische populatie per zone met leeftijd, geslacht en toegewezen leefstijl. Vervolgens worden OD-relaties gegenereerd, uitgesplitst naar leefstijl. Dit resulteert in zeven OD-matrices die gebruikt kunnen worden in het Fountain-model voor verplaatsingsprognoses.

6. Validatie. Validatie vindt plaats via twee casestudies. In Amsterdam is de uitkomst van de IPF Multizone vergeleken met bekende leefstijlverdelingen. In Utrecht is gekeken naar het effect van de synthetische populatie op de voorspelling van spitsmijdingsgedrag. In beide gevallen zijn verschillen tussen zones aantoonbaar gemaakt via MAPE-analyses (tussen 2% en 9%), wat wijst op plausibele variatie in leefstijlen.

7. Openbaarheid. Het document is een afstudeerscriptie en vrij beschikbaar. De gebruikte data (zoals de leefstijlverdeling van Anable) is open, evenals de gebruikte algoritmes (zoals IPF). De tool zelf lijkt niet openbaar beschikbaar als software, maar de beschrijving is voldoende gedetailleerd om de aanpak te reproduceren.

8. Toepasbaarheid voor SIVMO. De methode is relevant voor SIVMO vanwege de nadruk op gedragsdifferentiatie via leefstijlen in plaats van klassieke demografie. De gebruikte IPF-aanpak is transparant en aanpasbaar. De leefstijlbenadering past bij een bredere gedragsmatige benadering van mobiliteitsmodellen, al is de afhankelijkheid van slechts één brondataset (Anable) een zwakte. De methode is goed bruikbaar voor conceptuele verkenningen of beleidsanalyses waarbij leefstijlsegmentatie centraal staat, mits aanvullende validatie plaatsvindt.

Borysov, Stanislav S., Jeppe Rich & Francisco Camara Pereira. 2019.
"*How to Generate Micro-Agents? A Deep Generative Modeling Approach to Population Synthesis.*" Transportation Research Part C: Emerging Technologies. 106: 73–97. https://doi.org/10.1016/j.trc.2019.07.006 .
scalable_population_synthesis_rev1_clean_1.pdf.

Borysov, Stanislav S., Jeppe Rich & Francisco C. Pereira. 2019.
"Scalable Population Synthesis with Deep Generative Modeling." Transportation Research Part C: Emerging Technologies 106(2019), pp. 73-79. https://doi.org/10.1016/j.trc.2019.07.006. 180806910.pdf

Samenvatting

Beide artikelen zijn vrijwel identiek. Ze introduceren een nieuwe aanpak voor population synthesis met behulp van een Variational Autoencoder (VAE). In tegenstelling tot traditionele methoden zoals Iterative Proportional Fitting (IPF), Gibbs sampling of Bayesian Networks, biedt deze deep learning-methode betere mogelijkheden om hoge-dimensionale populaties realistisch te genereren. De methode is ontwikkeld om micro-agenten te synthetiseren die relevant zijn voor activity-based modellen, waarbij het doel is om statistisch realistische, maar unieke agents te creëren, zelfs buiten de grenzen van de trainingsdata.

De auteurs testen de VAE aan de hand van een grootschalige Deense reisdagboekdataset, waarin ze drie niveaus van attributensets onderscheiden: Basic (4 variabelen), Socio (21) en Extended (47). De VAE wordt vergeleken met Gibbs sampling en Bayesian Networks. Terwijl traditionele methoden beter presteren bij lage dimensionaliteit, blijken deze moeilijk schaalbaar wanneer het aantal variabelen stijgt. Gibbs sampling heeft bovendien de neiging om agents uit de trainingsdata exact te reproduceren, terwijl de VAE juist in staat is om nieuwe, plausibele combinaties te genereren.

Een belangrijk voordeel van de VAE is dat deze werkt met een latente ruimte, waarin data opnieuw worden opgebouwd. Dit maakt het mogelijk om variatie in de gegenereerde agents aan te brengen zonder directe kopieën te maken. Deze aanpak heeft ook implicaties voor privacybescherming: synthetische populaties kunnen worden gegenereerd zonder directe overeenkomsten met echte individuen. De validatie vindt plaats via metrics zoals SRMSE, correlaties en diversiteitsmaten, waarbij de VAE over het algemeen beter presteert bij hogere dimensies.

De auteurs concluderen dat deep generative models zoals de VAE veelbelovend zijn voor grootschalige toepassingen in transportmodellen, vooral in situaties waarin veel attributen vereist zijn of waarin privacy van belang is. Ze signaleren wel dat verdere ontwikkeling nodig is, bijvoorbeeld richting conditional VAE's of integratie met re-samplingmechanismen die gericht zijn op beleidsmatige targets. Ze positioneren de VAE als krachtige uitbreiding op de bestaande populatiesynthese-instrumenten.

Analyse

1. Methode. De studie introduceert een *Variational Autoencoder* (VAE), een deep generative model afkomstig uit de machine learning. Deze methode leert de volledige joint distribution van agentkenmerken via een encoder–decoder-architectuur met een latente representatie. Vergeleken met Gibbs sampling en Bayesian Networks laat de VAE vooral voordelen zien bij hoge dimensionaliteit. Er wordt ook uitgebreid ingegaan op de trainingsarchitectuur, optimalisatie, en validatiemethoden.

2. Inputdata. De methode is getraind op het Deense nationale reisdagboek (TU-dataset, 2006–2017) met ca. 146.000 respondenten en circa 1 miljoen trips. Dit vormt een rijke inputset van zowel numerieke als categorische variabelen. Er worden drie subsets onderscheiden met 4, 21 en 47 variabelen, afhankelijk van de complexiteit.

3. Kenmerken. De gegenereerde populaties bevatten micro-agents met een combinatie van socio-demografische kenmerken en mobiliteitsgegevens (zoals inkomen, gezinsgrootte, reisafstand, modaliteit, enz.). De VAE kan ook out-of-sample agents genereren, wat een belangrijke eigenschap is voor toekomstverkenningen.

4. Schaal. De focus ligt op de nationale schaal van Denemarken, met de mogelijkheid tot toepassing op kleinere zones vanwege de hoge resolutie. Er is geen expliciete tijdsdynamiek (zoals toekomstscenario's), maar de methode is geschikt voor simulatie van alternatieve populaties via re-sampling.

5. Output. De resultaten worden gepresenteerd in termen van distributie-vergelijkingen tussen synthetische en testdata: marginaal, bivariate en trivariate distributies, SRMSE, correlaties, PCA-plots en diversiteitsmaten. De presentatie is helder, met uitgebreide tabellen en figuren (zoals marginals per attribuut en projecties van de latente ruimte).

6. Validatie. De validatie is grondig uitgevoerd. Er wordt gewerkt met een aparte trainings- en testset, en diverse evaluatiematen zoals SRMSE, Cramér's V, nearest-sample afstand en PCA. De vergelijking met traditionele methoden (Gibbs, BN) maakt duidelijk waar de VAE sterker of zwakker presteert.

7. Openbaarheid. Het artikel is gepubliceerd in *Transportation Research Part C* en daarmee openbaar beschikbaar. De gebruikte code is gedeeld via GitHub (https://github.com/stasmix/popsynth). De gebruikte dataset (TU Denemarken) is niet vrij beschikbaar, maar toegankelijk voor onderzoeksdoeleinden.

8. Toepasbaarheid voor SIVMO. De aanpak is technisch geavanceerd en vooral interessant voor toepassingen waarbij hoge resolutie, heterogeniteit en privacy een rol spelen. De methode vereist echter substantiële expertise in deep learning en toegang tot GPU-hardware. Voor SIVMO is het relevant als aanvulling op bestaande methoden, vooral bij verkenning van toekomstscenario's of het modelleren van nieuwe populatiestructuren. Een vereenvoudigde VAE of samenwerking met universiteiten zou toepasbaarheid kunnen vergroten.

Borysov, Stanislav S., Jeppe Rich & Francisco C. Pereira. n.d.
"*Population Synthesis Meets Deep Generative Modelling*". Lyngby: Department of Management Engineering, Technical University of Denmark. 5380.pdf.

Samenvatting

Het artikel introduceert een nieuwe aanpak voor populatiesynthese in agent-based transportmodellen, gebaseerd op deep generative modelling. Traditionele methoden zoals IPF en MCMC schieten tekort bij hoge datadimensionaliteit. Ze zijn onvoldoende schaalbaar wanneer populaties met veel kenmerken nodig zijn, bijvoorbeeld bij fijnmazige ruimtelijke indelingen of combinatie van persoons- en huishoudkenmerken. De auteurs stellen voor om in plaats daarvan Variational Autoencoders (VAE's) te gebruiken, die in staat zijn om complexe verdelingen met veel variabelen te leren en synthetische populaties te genereren.

Een VAE is een neuraal netwerk dat data samenvat in een lagere-dimensionale ruimte (de 'latent space') en daaruit nieuwe combinaties genereert die statistisch overeenkomen met de oorspronkelijke data. Het model is volledig probabilistisch, waardoor het geschikt is voor toepassingen met onzekerheid, validatie via log-likelihood, en outlierdetectie. Deze eigenschappen bieden voordelen zoals privacybescherming (geen echte personen), datacompressie, imputatie van missende data, en analyse van patronen in de latente ruimte.

De methode is getest op de Deense Nationale Reisgegevens (TU), met ruim 23.000 records uit 2010. Hierbij zijn 37 kenmerken per persoon gesynthetiseerd (waaronder inkomen, reistijd en opleidingsniveau). De gegenereerde data benaderen de oorspronkelijke verdeling goed. In vergelijking met een eenvoudig basismodel toont de VAE betere prestatie op correlatie, RMSE en verklaringsgraad. Dit bewijst dat de methode ook in hoge dimensies robuust werkt, waar MCMC-methoden last hebben van combinatorische explosie.

Tot slot bespreken de auteurs de uitdaging van toekomstgerichte synthese. Hoewel directe integratie van margedoelen in een VAE nog onderwerp van onderzoek is, kunnen alternatieve methoden zoals quota- of herwegingstechnieken worden toegepast. Ook wordt uitbreiding naar huishoudens en activiteitenpatronen als interessante vervolgstap gezien.

Analyse

1. Methode. De studie introduceert een nieuwe aanpak voor populatiesynthese op basis van deep generative modelling, specifiek een Variational Autoencoder (VAE). De VAE is een neuraal netwerk dat een complexe multivariate verdeling leert en synthetische individuen genereert via sampling in een lage-dimensionale 'latent space'. De methode is volledig probabilistisch en schaalbaar. Er wordt ook kort gewezen op mogelijke uitbreiding richting toekomstscenario's met behulp van herwegingstechnieken.

2. Inputdata. De input bestaat uit data uit de Deense nationale reisgegevens (TU), waaronder 34 numerieke variabelen (zoals leeftijd, inkomen, reistijd) en 3 categorische (geslacht, opleiding, beroep). De input wordt gebruikt voor het trainen van het VAE-model. Er wordt geen gebruik gemaakt van marginaal beschikbare

randtotalen bij de synthese, al wordt dat als een toekomstig uitbreidingspad genoemd.

3. Populatiekenmerken. De gesynthetiseerde populatie bevat in totaal 37 kenmerken per individu, zowel sociaaleconomisch als mobiliteitsgerelateerd. De methode is in staat om hoge dimensionaliteit te verwerken en leert de onderliggende correlaties zonder expliciete afhankelijkheden te modelleren. De kenmerken zijn op persoonsniveau, zonder huishoudstructuur.

4. Schaal. De methode is getest op nationale schaal (Denemarken) met een dataset van 23.754 personen. De geografische spreiding binnen het land wordt niet als aparte variabele behandeld. De methode is echter generiek toepasbaar en schaalbaar naar grotere of fijnmazigere populaties.

5. Output. De output is een synthetische populatie die lijkt op de oorspronkelijke data, inclusief coherente combinaties van kenmerken. De VAE genereert individuen die statistisch plausibel zijn, maar niet herleidbaar naar bestaande personen. Er worden geen OD-matrices of verplaatsingsgegevens gegenereerd; de focus ligt op de populatie zelf.

6. Validatie. Validatie is uitgevoerd door de gegenereerde marginale en gezamenlijke verdelingen te vergelijken met de oorspronkelijke data. Statistische maatstaven (zoals correlatie, RMSE, R^2) tonen aan dat de VAE een betere fit levert dan een marginaal onafhankelijk basismodel. Validatie op beleidsrelevante gedragsuitkomsten ontbreekt echter.

7. Openbaarheid. De methode is gebaseerd op literatuur en algemeen beschikbare deep learning libraries. De paper zelf is vrij toegankelijk. Er is geen melding van open softwarecode of herbruikbare tool. De gebruikte TU-data is niet vrij beschikbaar.

8. Toepasbaarheid voor SIVMO. De methode is conceptueel vernieuwend en potentieel geschikt voor SIVMO, vooral vanwege de schaalbaarheid, het vermogen tot hoge dimensionaliteit en privacyvriendelijkheid. Voor praktische toepassing is echter verdere ontwikkeling nodig, met name rond integratie van marges, huishoudstructuur en lokale datasets. Voor scenarioanalyses met complexe populatiekenmerken is het een interessante kandidaat. Let op: de methode is probabilistisch. Reproductie is alleen mogelijk door een random seed vast te zetten tijdens de sampling.

Chapuis, Kevin & Patrick Taillandier. 2019.

A Brief Review of Synthetic Population Generation Practices in Agent-Based Social Simulation. Conference paper, September 2019.
https://www.researchgate.net/publication/335601121 .
A_review_of_synthetic_population_generation_process_in_social_simulation-4.pdf.

Samenvatting

Deze review bespreekt de gangbare praktijken in het genereren van synthetische populaties binnen agent-based social simulations (ABSS). Hoewel er diverse methoden zijn ontwikkeld, blijkt uit een analyse van artikelen in het Journal of Artificial Societies and Social Simulation (JASSS) dat deze technieken zelden daadwerkelijk worden toegepast. Veel modellen vertrouwen op eenvoudige, ad-hoc methoden om agentkenmerken te genereren, vaak zonder empirische basis. Dit staat haaks op de behoefte aan realistische, datagedreven simulaties.

De auteurs onderscheiden twee hoofdmethoden voor populatiesynthese: *synthetic reconstruction* (SR) en *combinatorial optimization* (CO). SR-methoden genereren populaties door karakteristieken te trekken op basis van bekende verdelingen, bijvoorbeeld via IPF of Bayesian netwerken. CO-methoden optimaliseren een steekproef ten opzichte van bekende marges via algoritmes als simulated annealing of genetische algoritmen. Hoewel krachtig, vereisen deze methoden vaak gedetailleerde en harmonische datasets, wat hun toepassing beperkt.

De analyse van 228 artikelen laat zien dat slechts een derde expliciet beschrijft hoe populaties zijn gegenereerd. De meeste modellen gebruiken eenvoudige randomisatie op basis van standaardverdelingen (zoals uniform of normaal), vaak gebaseerd op expertkennis of literatuur. Empirische databronnen worden zelden benut; slechts 16,5% van de modellen gebruikt steekproefdata. Gebruik van geavanceerde populatiesynthesemethoden zoals IPF of CO is uitzonderlijk.

De auteurs pleiten voor betere integratie van bestaande populatiesynthesetools in platforms zoals NetLogo of GAMA. Hiermee zou het gebruik van realistisch gesynthetiseerde populaties in ABSS kunnen toenemen, mits ook obstakels zoals dataverschillen en harmonisatie worden aangepakt. De studie sluit af met een oproep tot betere documentatie en openheid in modellering.

Analyse

1. Methode. De paper biedt een review van methoden voor het genereren van synthetische populaties in agent-based social simulations (ABSS). Twee hoofd-benaderingen worden onderscheiden: *synthetic reconstruction* (SR), waaronder IPF, Bayesian netwerken en MCMC vallen, en *combinatorial optimization* (CO), waarbij individuen uit een steekproef worden geselecteerd op basis van een fitnesfunctie. Deze methoden worden theoretisch uitgelegd, inclusief gebruikte algoritmen zoals simulated annealing, genetische algoritmen en hill climbing. De auteurs analyseren daarnaast welke methoden daadwerkelijk worden toegepast in praktijkstudies binnen JASSS (2014–2018).

2. Inputdata. De review toont dat inputdata in praktijkstudies vaak beperkt zijn of slecht gedocumenteerd. Ongeveer 28% van de modellen gebruikt enige vorm van

empirische data (zoals steekproeven of marges), maar in 72% van de gevallen ontbreekt een duidelijke beschrijving of wordt expertkennis gebruikt. Dit belemmert transparantie en reproduceerbaarheid. De auteurs merken op dat databases zoals IPUMS vrijwel nooit worden gebruikt.

3. Kenmerken. De besproken methoden genereren synthetische populaties met socio-demografische kenmerken, zowel op individueel als huishoudniveau. Voor SR-methoden geldt dat afhankelijkheden tussen kenmerken kunnen worden gemodelleerd (bv. via Bayesian netwerken). In de praktijk ligt de nadruk echter vaak op eenvoudige attributen (zoals leeftijd of opinie), en minder op complexe huishoudstructuren.

4. Schaal. De schaal van toepassing varieert sterk. In theorie zijn de besproken methoden inzetbaar op meerdere ruimtelijke niveaus, maar in de praktijk zijn veel modellen abstract en theoretisch. Empirische modellen zijn vaker op kleinere schaal (case studies). Tijdshorizon wordt zelden benoemd.

5. Output. De output van de besproken methoden zijn gesynthetiseerde individuen of huishoudens, die statistisch overeenkomen met bekende verdelingen. In de praktijk is de kwaliteit van deze output vaak onbekend, doordat validatie of beschrijving ontbreekt. Slechts 36% van de artikelen beschrijft expliciet hoe agentkenmerken zijn gegenereerd.

6. Validatie. Validatie is geen centraal onderwerp in dit reviewartikel. De auteurs constateren juist dat de validatie van gesynthetiseerde populaties in ABSS vaak ontbreekt of niet wordt beschreven. Fitnesfuncties zoals SRMSE en TAE worden genoemd, maar hun toepassing in de praktijk blijkt zeer beperkt.

7. Openbaarheid. Het artikel bespreekt publiek beschikbare methoden zoals IPF en Bayesian netwerken, en verwijst naar eerdere literatuur en open data-instrumenten. De auteurs signaleren echter dat slechts een klein deel van de studies toegang geeft tot broncode of ODD-documentatie. Herhaalbaarheid van simulaties is daardoor beperkt.

8. Toepasbaarheid voor SIVMO. De paper is relevant voor SIVMO in de zin dat het inzicht biedt in de kloof tussen methodologische mogelijkheden en praktische toepassingen van populatiesynthese in agent-based modellen. De classificatie van methoden en de analyse van toepassing in de praktijk zijn bruikbaar voor het beoordelen van transparantie, reproduceerbaarheid en datagebruik binnen Nederlandse modellen. Vooral de afwezigheid van standaardmethoden en datagebruik in veel praktijkmodellen is een belangrijke observatie.

Choupani, Abdoul-Ahad & Amir Reza Mamdoohi. 2016.
"Population Synthesis Using Iterative Proportional Fitting (IPF): A Review and Future Research." Transportation Research Procedia 17: 223–233.
https://doi.org/10.1016/j.trpro.2016.11.078. 1-s20-S2352146516306925-main.pdf.
Population_Synthesis_Using_Iterative_Proportional_ (1).pdf & 1-s20-S2352146516306925-main.pdf

Samenvatting

Deze reviewstudie onderzoekt het gebruik van het Iterative Proportional Fitting (IPF) algoritme binnen populatiesynthese voor activity-based models. IPF is een wijdverbreide methode vanwege zijn eenvoud, snelheid, beperkte databronvereisten en gegarandeerde convergentie. De studie behandelt hoe IPF werkt, onderscheid maakt tussen 'fitting' en 'allocatie', en welke veelvoorkomende problemen optreden, zoals de noodzaak tot integerconversie en het omgaan met lege cellen. Ook bespreekt het artikel verschillende implementaties en softwaretools die IPF toepassen, en hoe synthesefouten ontstaan afhankelijk van steekproefgrootte en ruimtelijke aggregatie.

Het artikel geeft een uitgebreide typologie van bestaande IPF-gebaseerde synthesizers en behandelt hun prestaties op aspecten zoals validatie (intern vs extern), categorisatiestrategieën, schaal van toepassing en de keuze van controlevariabelen. De auteurs benadrukken dat de meeste validaties intern zijn en dat externe validatie vaak ontbreekt. Dit leidt tot een overschatting van de betrouwbaarheid. Daarnaast toont de studie dat heterogeniteit tussen zones, zoals in voertuigeigendom, foutmarges beïnvloedt. Tolerantiegrenzen van foutmarges worden besproken, net als methoden om daarmee om te gaan.

Een groot deel van de analyse richt zich op problemen met integerconversie (afronding van fracties naar gehele aantallen) en zero-cell problematiek. Er worden methoden besproken zoals category aggregation, tweaking en het overnemen van waarden uit andere gebieden. Deze hebben elk voor- en nadelen, maar geen biedt een volledig onbevooroordeelde oplossing. De auteurs pleiten daarom voor controlled rounding-methoden gebaseerd op optimalisatie.

Tot slot signaleert het artikel een opkomende trend: simulatiegebaseerde synthese (bijv. Gibbs sampling). Deze methode omzeilt het fittingproces van IPF en kan beter omgaan met hoge dimensionaliteit en sparsity. De auteurs concluderen dat toekomstige verbeteringen in populatiesynthese vooral liggen in robuuste validatiekaders, het oplossen van integer- en zero-cellproblemen, en het integreren van nieuwe methoden zoals simulatiebenaderingen.

Analyse

1. Methode. De studie is een literatuurreview gericht op het IPF-algoritme in populatiesynthese. De auteurs geven een gedetailleerde beschrijving van het tweeledige IPF-proces (fitting en allocatie), en behandelen varianten zoals directe en indirecte integerconversie, single- vs. multi-level synthese, en categorisatiemethoden. Het artikel presenteert een gestructureerd overzicht van IPF-toepassingen, hun tekortkomingen (o.a. zero cells, integerisatie, validatie) en alternatieven, zoals simulatiegebaseerde synthese (bijv. Gibbs sampling).

2. Inputdata. De studie is conceptueel. Er worden geen eigen datasets gebruikt, maar bestaande toepassingen van synthesetools worden geïnventariseerd (TRANSIMS, PopGen, CEMDAP e.a.). De auteurs gaan in op welke variabelen (bijv. inkomen, leeftijd) vaak gebruikt worden en welke categorisatieschema's worden gehanteerd. Verschillen tussen huishoud- en persoonsniveau worden besproken.

3. Populatiekenmerken. IPF wordt doorgaans toegepast op socio-demografische kenmerken zoals huishoudenstype, inkomen, leeftijd en gezinsgrootte. De auteurs beschrijven hoe populaties op één of twee niveaus (huishouden en persoon) kunnen worden gesynthetiseerd, en hoe problemen ontstaan bij het combineren van meerdere variabelen of kleine celwaarden (sparsity). Er wordt gepleit voor combinatie van attributen over meerdere niveaus met aangepaste IPF- of gewichtsmodellen.

4. Schaal. De bespreking beslaat toepassingen op zowel stads-, regio- als landsniveau. De auteurs tonen aan dat foutmarges toenemen bij kleinere zones of grotere heterogeniteit. Synthesetools proberen dat vaak te vermijden door grotere eenheden te gebruiken of aggregatie toe te passen.

5. Output. De output van IPF-methodes bestaat uit synthetische populaties (tabelvormig), waarin huishoudens en personen zijn toegewezen aan zones met bijbehorende kenmerken. De auteurs beschrijven zowel fractionele tabellen (na fitting) als discrete agenten (na allocatie). Er is aandacht voor de gevolgen van afronding en selectiemethoden, zoals sampling with replacement en probabilistische selectie op basis van gewichten.

6. Validatie. De auteurs onderscheiden interne en externe validatie, en bespreken de valkuil dat interne validatie (op gebruikte controlevariabelen) meestal betere resultaten oplevert dan externe validatie (op onafhankelijke variabelen). Een kritiek-punt is dat fitting, integerisatie en selectie zelden afzonderlijk worden gevalideerd. De auteurs pleiten voor een systematisch validatiekader die deze stappen toetst.

7. Openbaarheid. De besproken synthesetools zijn deels open, deels institutioneel ontwikkeld. Hoewel het artikel geen eigen software presenteert, bevat het een overzichtstabel van 15 bestaande synthesizers (ARC, ILUTE, PopGen, PopSynWin, TRANSIMS, FSUTMS, CEMDAP, ALBATROSS, SimBritain, MORPC, TRESIS, OREGON2, SFCTA, METRO en BNY) met verwijzingen naar de ontwikkelaars en publicaties.

8. Toepasbaarheid voor SIVMO.
Deze studie is vooral relevant voor SIVMO als overzichtswerk. Het biedt een grondige analyse van de zwakke plekken van IPF bij hoge dimensies, kleine zones en validatievraagstukken. De nadruk op zero-cells, integerisatie en schaalproblemen is zeer bruikbaar bij het ontwerpen van een robuust syntheseproces. Ook de bespreking van simulatiegebaseerde alternatieven (zoals Gibbs sampling) sluit aan bij modernere, flexibelere benaderingen. Hoewel geen directe tool wordt geleverd, is dit een sleutelpublicatie voor methodologische onderbouwing.

Feng, Lewen & Md. Kamruzzaman. 2023.
"*Comparing Major Population Synthesis Techniques: A Case Study in Monash, Victoria*." Paper presented at the Australasian Transport Research Forum 2023, Perth, Australia, November 29 – December 1, 2023. ATRF_2023_Paper_114.pdf.

Samenvatting

De paper van Feng en Kamruzzaman (2023) vergelijkt drie veelgebruikte technieken voor populatiesynthese: *Iterative Proportional Fitting* (IPF), *Iterative Proportional Updating* (IPU) en *Simulated Annealing* (SA). Deze technieken worden toegepast in agent-based verkeersmodellen, waarin individuele gedragsdata nodig zijn. Omdat volledige populatiedata moeilijk te verkrijgen zijn, worden synthetische populaties gegenereerd uit kleinere steekproeven. De auteurs benoemen de fitting- en generatiefase van het syntheseproces en beschrijven hoe IPF, IPU en SA verschillende benaderingen gebruiken om steekproeven te herwegen en volledige populaties te creëren.

De case study is uitgevoerd in Monash, een regio nabij Melbourne. Als brondata diende de VISTA-enquête (2012–2018) met 1.043 huishoudens en 2.816 personen. Deze zijn aangevuld met totalen uit de Australische volkstelling van 2016. Slechts enkele categorieën kwamen overeen tussen beide datasets, waardoor selectie van compatibele variabelen noodzakelijk was. Voor elk van de drie methoden werd een identiek invoerbestand gebruikt, verwerkt via het *simPop* R-pakket. Voor SA was extra verwerking nodig met behulp van multivariate verdelingen.

De resultaten tonen dat IPF het beste presteert op huishoudniveau, terwijl IPU beter scoort op persoonsniveau. SA zit daar tussenin, maar kent afwijkingen in de verdeling van huishoudgrootte. De IPF-methode convergeert snel, omdat ze slechts één constraintniveau gebruikt. IPU en SA hanteren dubbele constraints, wat leidt tot grotere afwijkingen op huishoudniveau maar betere persoonsgegevens. Voor alle methoden zijn marginale verdelingen redelijk accuraat, met uitzondering van leeftijdscategorieën.

De auteurs concluderen dat elke methode zijn sterke en zwakke punten heeft. IPF is nauwkeurig op huishoudniveau, IPU op persoonsniveau. De geschiktheid hangt af van het type toepassing in verkeersmodellen. Vergelijking van synthetische reisinformatie met werkelijke gegevens vormt een volgende stap in het onderzoek. Deze studie levert waardevolle inzichten voor de keuze van synthesemethoden in mobiliteitsanalyses.

Analyse

1. Methode. De studie vergelijkt drie populatiesynthesetechnieken: *Iterative Proportional Fitting* (IPF), *Iterative Proportional Updating* (IPU) en *Simulated Annealing* (SA). Elk van deze technieken bestaat uit twee stappen: een fitting-fase waarin wegingen worden aangepast op basis van marges, en een generatiefase waarin synthetische huishoudens worden gecreëerd. Voor IPF en IPU gebeurt de fitting via de ipu() functie in het *simPop* R-pakket; SA maakt gebruik van calibPop() met joint marges. De output wordt afgerond naar gehele aantallen (IPF/IPU) of direct

gegenereerd (SA). De prestaties worden beoordeeld via absolute foutpercentages ten opzichte van censusdata.

2. Inputdata. De input bestaat uit microdata van VISTA (2012–2018) met 1.043 huishoudens en 2.816 personen in Monash (Victoria), gecombineerd met control totals uit de Australische volkstelling van 2016. Slechts een beperkt aantal categorieën komt overeen tussen de bronnen (bijv. huishoudgrootte, voertuigbezit, geslacht), waardoor de analyse zich noodgedwongen op een subset van variabelen moest richten. SA vereist multivariate marges, afgeleid uit de Census.

3. Populatiekenmerken. De synthetische populatie bevat basiskenmerken op huishoud- en persoonsniveau: o.a. huishoudgrootte, type woning, aantal voertuigen, leeftijd en geslacht. IPF houdt alleen rekening met huishoudkenmerken, IPU en SA met beide niveaus. Er is geen sprake van gedragstoewijzing of kenmerken zoals opleiding of mobiliteitspatronen. De studie heeft tot doel dit later te analyseren.

4. Schaal. De toepassing betreft het SA3-gebied Monash (30.000–130.000 inwoners), een subregio binnen Melbourne. Dit maakt de analyse representatief voor een stedelijk gebied met voldoende steekproefomvang, maar beperkt tot één geografische eenheid. De methode is herbruikbaar voor andere gebieden mits vergelijkbare data beschikbaar zijn.

5. Output. Elke techniek levert een synthetische populatie op met eenheden op huishoud- en persoonsniveau. De evaluatie van de output gebeurt door vergelijking van marginale verdelingen met censusdata. IPF presteert beter op huishoudniveaus (fout ~0.02%), IPU beter op persoonsniveau (fout ~3.7%), terwijl SA ertussenin zit. Voor leeftijd zijn grotere afwijkingen geconstateerd.

6. Validatie. De validatie is intern: synthetische verdelingen worden vergeleken met de bekende totalen van de census. Er is geen externe validatie (bijv. met ongebruikte variabelen of reisinformatie), hoewel dat wel als toekomstig werk wordt genoemd. De foutmaten zijn gebaseerd op absolute afwijkingen per attribuutcategorie.

7. Openbaarheid. De gebruikte VISTA-data zijn niet publiek beschikbaar, maar de censusdata wel. De toegepaste methoden zijn open source via het *simPop* R-pakket (https://github.com/statistikat/simPop). De implementatie is reproduceerbaar, mits toegang tot VISTA. De studie beschrijft de werkwijze voldoende transparant, maar de gebruikte scripts zijn niet gedeeld.

8. Toepasbaarheid voor SIVMO. De studie is relevant voor SIVMO vanwege de directe vergelijking van meerdere algoritmen. Vooral de implicaties voor activity-based modelling zijn interessant. De eenvoudige dataselectie (weinig variabelen) is een beperking, maar de methodevergelijking en toepassing op een stedelijk gebied bieden praktische aanknopingspunten voor scenarioverkenning. Uitbreiding naar gedrag (zoals mobiliteit) is nog nodig om de bruikbaarheid volledig in te schatten.

Fournier, Nicholas, Eleni Christofa, Arun Prakash Akkinepally & Carlos Lima Azevedo. 2018.
"*An Integration of Population Synthesis Methods for Agent-Based Microsimulation*." Extended abstract submitted to the 97th Annual Meeting of the Transportation Research Board (TRB), Washington, D.C. CLA_Population_TRB_2018_extabs.pdf.

Samenvatting

Fournier et al. (2018) presenteren een geïntegreerd raamwerk voor populatiesynthese ten behoeve van agent-based microsimulatie. De studie erkent dat er geen universeel superieure methode bestaat, gezien verschillen in datatoegang en ruimtelijke schaal. Daarom combineren de auteurs verschillende technieken in één proces. De aanpak gebruikt Amerikaanse Censusdata (waaronder PUMS) en bestaat uit vijf stappen: seeding, Iterative Proportional Fitting (IPF), integerisatie via de TRS-methode (Truncate-Replicate-Sample), Iterative Proportional Updating (IPU) en Monte Carlo-sampling. Het doel is het genereren van een synthetische populatie die geschikt is voor grootschalige simulaties.

Het raamwerk wordt toegepast op de Greater Boston Area (GBA), een regio met 4,6 miljoen mensen en 1,7 miljoen huishoudens. De PUMS-data wordt per census tract geschaald, met fallback-methodes om zero cells te vermijden zonder structurele informatie te verliezen. Vervolgens worden IPF-resultaten integer gemaakt met TRS, zodat zeldzame huishoudens behouden blijven. De IPU-stap wordt geoptimaliseerd via C++-code en gebruik van sparse matrices. Monte Carlo-sampling genereert uiteindelijk een synthetische populatie op tractniveau, waarbij per tract de best passende steekproef wordt geselecteerd.

Validatie vindt plaats op twee niveaus: marginaal (per regio en tract) en op microdata-niveau (typen huishoudens en personen). De resultaten zijn goed op marginaal niveau (NRMSE < 0.5), maar minder op microdata-niveau (NRMSE > 6). Dit wijst op hoge algemene nauwkeurigheid, maar beperkte representatie van individuele typen. Het model werkt snel (onder twee uur), behoudt veel detail en is overdraagbaar naar andere regio's in de VS.

De studie concludeert dat de combinatie van bestaande en innovatieve technieken tot efficiënte, schaalbare populatiesynthese leidt. Verbeterpunten liggen bij gedragsdata-integratie en verdere optimalisatie van IPU.

Analyse

1. Methode. De auteurs combineren vijf stappen in een geïntegreerd synthesemodel: (1) seeding via PUMS-data op tractniveau, (2) *Iterative Proportional Fitting* (IPF), (3) integerisatie via de TRS-methode (Truncate, Replicate, Sample), (4) *Iterative Proportional Updating* (IPU), en (5) Monte Carlo-sampling. Door IPF en IPU te combineren met integerisatie en sparse matrixgebruik, wordt het synthesemodel zowel nauwkeurig als efficiënt. De IPU-stap is deels herschreven in C++ voor snelheid.

2. Inputdata. De input bestaat uit microdata van de Amerikaanse *Public Use Microdata Sample* (PUMS), en control totals uit de US Census (2010). Er zijn 5 huishoudkenmerken en 8 persoonskenmerken gebruikt, waaronder inkomen, leeftijd,

voertuigen, ras, schoolinschrijving, beroep en reistijd. De data zijn vrij beschikbaar en toepasbaar voor de hele VS, mits PUMS en censusgegevens voor de regio beschikbaar zijn.

3. Populatiekenmerken. De gesynthetiseerde populatie bevat gedetailleerde kenmerken per huishouden en persoon. Zowel socio-demografische als gedragsgerelateerde variabelen worden meegenomen. Het model is geschikt voor microsimulatie op huishoudniveau met individuen die onderling gekoppeld zijn aan huishoudens, inclusief een breed spectrum aan variabelen.

4. Schaal. De toepassing is uitgevoerd voor de Greater Boston Area met 960 census tracts, 4,6 miljoen personen en 1,7 miljoen huishoudens. Het model is echter generiek en overdraagbaar naar andere regio's in de VS waar PUMS beschikbaar is. De schaal varieert van tractniveau (lokale analyse) tot regionaal niveau.

5. Output. De output is een synthetische populatie op huishoud- en persoonsniveau, met behoud van microstructuren en onderlinge samenhang. Er wordt een volledig agent-bestand gecreëerd dat direct bruikbaar is in agent-based modellen. Validatie gebeurt op marginaal en microdataprofielniveau, met cijfers per tract en voor de gehele regio.

6. Validatie. Validatie wordt uitgevoerd via NRMSE en regressieanalyse (R^2 en slope). Op marginaal niveau zijn de resultaten goed (NRMSE 0.27–0.47, $R^2 > 0.9$), maar op microdata-niveau aanzienlijk minder (NRMSE ~6.6–6.8, R^2 0.43–0.69). Dit wijst op goede matching van totalen, maar matige nauwkeurigheid van combinaties van kenmerken.

7. Openbaarheid. Alle gebruikte data zijn afkomstig van open bronnen (PUMS, Census). De methodologische beschrijving is gedetailleerd, inclusief verwijzing naar gebruikte softwarepakketten (zoals het *Matrix* R-pakket). De IPU-code is herschreven in C++, maar die implementatie is niet openbaar gedeeld. De werkwijze is wel reproduceerbaar met vergelijkbare tools.

8. Toepasbaarheid voor SIVMO. De studie is relevant voor SIVMO vanwege de combinatie van nauwkeurigheid, schaalbaarheid en reproduceerbaarheid. De aanpak is vooral geschikt bij beschikbaarheid van microdata en marges, en biedt oplossingen voor integerisatie en zero cells. Bevat geen gedragscomponenten zoals mobiliteit, maar biedt een robuuste basis voor populatie-initiatie in agent-based modellen.

Hafezi, Mohammad Hesam & Muhammad Ahsanul Habib. 2014. *"Synthesizing Population for Microsimulation-Based Integrated Transport Models Using Atlantic Canada Micro-Data."* Procedia Computer Science 37: 410–415. https://doi.org/10.1016/j.procs.2014.08.061. 1-s20-S1877050914010266-main.pdf

Samenvatting

Hafezi en Habib (2014) presenteren een synthetisch populatiekader voor agent-based microsimulatie in Atlantic Canada, als bouwsteen voor een geïntegreerd transport-, landgebruik- en milieumodel. Omdat microdata vaak niet beschikbaar zijn vanwege privacy of ontoegankelijkheid, is populatiesynthese noodzakelijk om een realistisch gedesaggregeerd databestand te genereren. De auteurs hanteren hiervoor de *Fitness-Based Synthesis* (FBS)-methode, die zowel huishoud- als persoonskenmerken tegelijk kan synthetiseren. FBS vermijdt beperkingen van traditionele IPF-methoden, zoals zero cells en afrondingsproblemen, en biedt tegelijk goede schaalbaarheid en rekenefficiëntie.

De methode gebruikt microdata uit de Public Use Microdata File (PUMF) van de Canadese volkstelling van 2006 en aggregaatdata van dezelfde bron. In elke iteratie selecteert het algoritme huishoudens op basis van hun 'fitness'-waarde, die aangeeft hoe goed ze bijdragen aan het benaderen van de control tables. De procedure stopt als er geen positieve fitnesswaarden meer zijn. De implementatie is uitgevoerd in MATLAB, gebruikmakend van sparse matrixtechnieken om geheugenverbruik en rekentijd te reduceren. Een GUI-prototype is meegeleverd als hulpmiddel bij toepassing.

Voor de validatie zijn twee varianten getest: met alleen huishoudcontroles (één niveau) en met gecombineerde huishoud- en persoonscontroles (twee niveaus). De variant met twee niveaus levert significant betere resultaten voor persoonskenmerken (bijv. geslacht, leeftijd, etniciteit), maar iets slechtere voor huishoudvariabelen. Errorpercentages dalen van ruim 10% naar circa 2% voor persoonskenmerken, wat de meerwaarde van multilevel controle aantoont. Attributen met minder categorieën geven bovendien een betere fit.

De studie concludeert dat FBS effectief is voor het synthetiseren van een realistische populatie met meerdere controlelagen. De methode is schaalbaar en efficiënt en vormt een bruikbare basis voor microsimulatie op regionaal niveau. Verdere stappen zijn toepassing op 100% populatie en verfijning naar gemeentelijk niveau zoals Halifax.

Analyse

1. Methode. De studie gebruikt *Fitness-Based Synthesis* (FBS), een iteratieve methode die huishoudens selecteert op basis van een berekende fitnesswaarde. Deze waarde drukt uit hoe goed een huishouden past bij de verschilmatrix tussen de control tables en de actuele tellingen. In tegenstelling tot IPF en IPU vereist FBS geen joint multivariate distributie en werkt het met integer gewichten. Het algoritme is geïmplementeerd in MATLAB, met gebruik van sparse matrices en een GUI.

2. Inputdata. De input bestaat uit microdata uit de *Public Use Microdata File* (PUMF) van de Canadese volkstelling van 2006 (steekproef van 1%), en bijbehorende aggregaatdata van Statistics Canada. Het betreft zowel huishoud- als persoons-

kenmerken. Er zijn 9 attributen opgenomen, waaronder leeftijd, geslacht, etniciteit, huishoudgrootte en woningtype. Vanwege databeperkingen is gekozen voor Atlantic Canada in plaats van alleen Halifax.

3. Populatiekenmerken. De synthetische populatie bevat huishoudkenmerken (grootte, inkomen, woningtype, eigendom) en persoonskenmerken (leeftijd, geslacht, etniciteit, immigratiestatus, burgerschap). Hiermee wordt een gedetailleerde agent-populatie gegenereerd, geschikt voor integratie in microsimulatiemodellen. Gedragskenmerken zoals mobiliteit of activiteit zijn niet inbegrepen, maar het model is voorbereid op verdere toepassing.

4. Schaal. De toepassing is regionaal: Atlantic Canada (in plaats van oorspronkelijk geplande Halifax). De schaal is dus beperkt tot een landsdeel, maar het model is schaalbaar naar grotere of kleinere gebieden mits data beschikbaar zijn. De implementatie is getest op een dataset van 9.142 huishoudens en 4.108 personen (1%-steekproef).

5. Output. De output is een synthetische populatie op individueel en huishoudniveau, met integer gewogen eenheden. Er is onderscheid gemaakt tussen populaties gesynthetiseerd met één en twee controlelagen. De versie met twee niveaus toont aanzienlijk betere overeenkomst op persoonskenmerken, terwijl huishoudkenmerken iets minder precies zijn. Validatie gebeurt via errorpercentages per attribuut.

6. Validatie. De validatie bestaat uit errorpercentages per attribuut en goodness-of-fit analyses met trendlijnen en R^2. De methode met twee controlelagen presteert beter voor persoonskenmerken (errors <5%) dan de éénlaagse variant (errors >10%). De validatie is intern; er is geen vergelijking met externe gedragsdata of mobiliteits-uitkomsten.

7. Openbaarheid. Het artikel en de gebruikte data zijn vrij toegankelijk. De PUMF en Censusdata zijn publiek via Statistics Canada. De FBS-methode is goed beschreven, inclusief formules. De gebruikte software (MATLAB) en algoritmische aanpak zijn reproduceerbaar, hoewel de daadwerkelijke broncode niet is gedeeld.

8. Toepasbaarheid voor SIVMO. De methode is relevant voor SIVMO vanwege de mogelijkheid om huishoud- én persoonskenmerken gelijktijdig te synthetiseren zonder joint distributions. De aanpak is transparant, integer en schaalbaar. Het ontbreken van gedragsdata en mobiliteitsvariabelen maakt het minder geschikt voor directe beleidsanalyses, maar als basispopulatie binnen een agent-based mobiliteitsmodel is FBS goed bruikbaar.

Harland, Kirk, Alison Heppenstall, Dianna Smith & Mark Birkin. 2012.
Creating realistic synthetic populations at varying spatial scales: A comparative critique of population synthesis techniques. Journal of Artificial Societies and Social Simulation, 15 (1). 1. ISSN 1460-7425 . https://doi.org/10.18564/jasss.1909 . Creating Realistic Synthetic Populations_published.pdf

Samenvatting

Dit artikel onderzoekt drie gangbare technieken voor populatiesynthese: deterministisch herwegen, conditionele waarschijnlijkheden en simulated annealing. De auteurs vergelijken deze methoden op hun vermogen om realistische synthetische populaties te genereren op verschillende geografische schaalniveaus. Ze testen de methoden met behulp van gegevens uit de Britse volkstelling van 2001 in het district Leeds, waarbij ze prestaties beoordelen op basis van totale foutmaten en het vermogen om bekende en onbekende kenmerken te reproduceren.

Deterministisch herwegen blijkt snel en eenvoudig te implementeren, maar is gevoelig voor de volgorde van constraints en geeft vaak slechtere resultaten bij kleinere geografische eenheden. De methode neigt tot het gladstrijken naar het gemiddelde van de steekproef, waardoor plaatselijke afwijkingen verloren gaan. Conditionele waarschijnlijkheden gebruiken Monte Carlo-sampling en bieden betere flexibiliteit, maar vereisen meer voorbereiding en kunnen moeite hebben bij veel constraints of beperkte steekproefdata.

Simulated annealing presteert het best over de meeste evaluaties. Het algoritme gebruikt een geavanceerde optimalisatieaanpak waarbij ook verslechteringen tijdelijk worden toegestaan om uit lokale minima te ontsnappen. Dit leidt tot een betere benadering van echte populaties, vooral op fijnmazige geografische schaal. Het nadeel is een hogere rekenlast.

De studie concludeert dat geen enkele onderzochte methode superieur is. Wel levert simulated annealing doorgaans de meest nauwkeurige resultaten, terwijl deterministisch herwegen goed toepasbaar blijft voor toepassingen met beperkte middelen of wanneer de focus ligt op specifieke uitkomsten. De combinatie van realistische synthese en verdere koppeling aan agent-based modellen wordt als veelbelovend gezien.

Analyse

1. Methode. Het artikel vergelijkt drie technieken voor statische populatiesynthese: deterministisch herwegen, conditionele waarschijnlijkheden en simulated annealing. Elke methode wordt toegepast om een synthetische populatie te genereren, waarbij de synthetische resultaten worden vergeleken met volkstellingsdata. De analyse omvat zowel univariate als multivariate vergelijkingen en evaluaties van niet-geconstrueerde variabelen. Er is sprake van een systematische en experimentele opzet met nauwkeurige foutmaten.

2. Inputdata. De studie maakt gebruik van het 2001 Census microdatabestand voor Leeds (SAMs) als steekproef, en geaggregeerde volkstellingstabellen als constraints (zoals leeftijd, geslacht, etniciteit, opleiding). Deze input is typisch voor Britse

toepassingen, maar vergelijkbare bronnen zijn in Nederland aanwezig (zoals CBS microdata en wijkstatistieken).

3. Kenmerken. De gegenereerde populaties bevatten persoonskenmerken zoals leeftijd, geslacht, etniciteit, opleidingsniveau en sociaaleconomische status. De populaties zijn op individueel niveau synthetisch gereconstrueerd en afgestemd op kleine geografische eenheden (output areas van ~300 personen).

4. Schaal. De toepassing richt zich op één regio (Leeds MDA) en drie geografische niveaus: OA (±300 personen), LLSOA en MLSOA. De methode is statisch, d.w.z. zonder tijdsdimensie. De schaal is goed vergelijkbaar met Nederlandse wijk- of buurtstatistiek.

5. Output. De output bestaat uit synthetische individuen met attributen, inclusief populaties per gebiedseenheid. Er worden ook inter-attribuutrelaties geëvalueerd. Daarnaast worden externe variabelen (niet in constraints) gebruikt om generaliseerbaarheid te toetsen.

6. Validatie. De studie hanteert kwantitatieve validatiemaatstaven, met name Total Absolute Error (TAE), Classificatiefout (CE) en %CE. Er is sprake van systematische vergelijking per methode, attribuut en schaalniveau. Simulated annealing blijkt het best te presteren, vooral op fijnmazige schaal.

7. Openbaarheid. De methoden zelf zijn openbaar en goed beschreven, inclusief pseudocode en formules (appendices A–D). De gebruikte data (Census en SAMs) zijn echter alleen toegankelijk voor geautoriseerde onderzoekers. De paper zelf is vrij beschikbaar.

8. Toepasbaarheid voor SIVMO. De studie biedt waardevolle inzichten voor SIVMO. Simulated annealing blijkt robuust en geschikt voor fijnmazige populatiesynthese. De evaluaties zijn herhaalbaar voor Nederlandse context, mits voldoende microdata beschikbaar zijn. De methodologische vergelijking is direct relevant voor afwegingen tussen eenvoud, snelheid en nauwkeurigheid.

Hörl, Sebastian & Miloš Balać. 2020.
"Open Data Travel Demand Synthesis for Agent-Based Transport Simulation: A Case Study of Paris and Île-de-France". Working Paper, Transport and Mobility Laboratory, EPFL. ab1499.pdf.

Samenvatting

In deze studie beschrijven Hörl en Balac hoe met uitsluitend open data een synthetische populatie en daaraan gekoppeld een verplaatsingsdagschema kan worden opgebouwd voor toepassing in een agent-based model (ABM) van de regio Île-de-France, inclusief Parijs. Dit wordt gedaan ter voorbereiding op een grootschalige MATSim-simulatie.

De auteurs combineren verschillende databronnen: censusgegevens (INSEE), verplaatsingsgegevens (Enquête Globale Transport, EGT), geografische informatie (OpenStreetMap) en reisinformatie (GTFS). Ze hanteren een rule-based approach voor het toekennen van activiteitenketens, vervoermiddelen en tijdstippen, met probabilistische sampling uit de EGT voor specifieke gedragsverdelingen. De populatie zelf wordt gegenereerd door sampling uit censusdata met aanvullende matchingregels. Huishoudenstructuren worden expliciet opgebouwd en gekoppeld aan individuele activiteitenprofielen.

De nadruk ligt op de reproduceerbaarheid en transparantie van het proces, mede dankzij het gebruik van volledig publieke en documenteerbare bronnen. Het resultaat is een volledige populatie van 12,1 miljoen agents, waarvan 10,1 miljoen actieve (d.w.z. met activiteiten buiten de woning). De populatie omvat meer dan 30 miljoen activiteiten en verplaatsingen op een gemiddelde dag. Het eindproduct wordt gebruikt als input voor een open MATSim-model van de regio.

Hoewel geen uitgebreide validatie plaatsvindt in dit paper zelf, verwijzen de auteurs naar een begeleidend artikel dat de realistische werking van het model beschrijft. Het doel van deze publicatie is vooral het delen van het opbouwproces en de bruikbaarheid van open data voor synthetische vraagmodellering.

Analyse

1. Methode. Men gebruikt een combinatie van methoden voor populatiesynthese en mobiliteitsgeneratie: probabilistische toekenning van persoons- en huishoud-kenmerken op basis van INSEE-statistieken, gevolgd door gedragstoekenning (activiteiten en verplaatsingen) via regels en distributies ontleend aan de Enquête Globale de Transport (EGT). Activiteitenlocaties worden gesimuleerd met behulp van ruimtegebruikdata en infrastructuurinformatie (OpenStreetMap en GTFS). Voor de verplaatsingsmodellen maken ze gebruik van MATSim, waarin het gegenereerde gedrag wordt doorgerekend. De populatie wordt gegenereerd voor de gehele regio Île-de-France.

2. Inputdata. Alle data zijn open en publiek beschikbaar: INSEE (statistiekbureau) levert huishoud- en persoonskenmerken; EGT biedt inzichten in activiteitenpatronen, modal split en ritkenmerken; OpenStreetMap en GTFS geven input voor netwerken en dienstregeling. De gebruikte data zijn actueel en representatief voor de regio Parijs

3. Kenmerken. De synthetische populatie bevat persoonskenmerken (leeftijd, geslacht, werkstatus, opleiding) en huishoudkenmerken (samenstelling, locatie). Gedrag wordt toegevoegd in de vorm van wekelijkse activiteitenschema's (inclusief typen, tijden, duur en locatie) en verplaatsingsketens. Dit maakt de populatie geschikt voor gedragsgebaseerde simulatie. Er is geen koppeling met attitudes of voorkeuren.

4. Schaal. De toepassing is regionaal (Île-de-France, ca. 12 miljoen inwoners), met een hoge resolutie: gemeenteniveau voor woonlocaties en gedetailleerde netwerken voor verplaatsingen. De temporele schaal betreft een werkdag (dynamisch gedrag voor 24 uur), geschikt voor dag-simulaties en beleidsinterventies op korte termijn.

5. Output. Een volledige synthetische populatie inclusief activiteitenketens, in MATSim-formaat. Elk individu heeft een dagelijks plan met meerdere activiteiten en trips. Output wordt gebruikt voor grootschalige ABM-simulaties.

6. Validatie. De resultaten worden gevalideerd aan de hand van EGT-data. Hierbij worden distributies van ritafstanden, vertrek- en aankomsttijden, activiteitenduur, en modal split vergeleken met observaties. De fit is over het algemeen goed, al is OV-gebruik iets onderschat. Ook netwerkbelasting wordt kwalitatief vergeleken met waargenomen congestiepatronen.

7. Openbaarheid. Alle gebruikte databronnen zijn open. De code en scripts zijn beschikbaar via GitHub: https://github.com/eqasim-org/ile-de-france . Het project is volledig reproduceerbaar. Licentie-informatie is niet gespecificeerd, maar de aanpak is bedoeld als referentie-implementatie.

8. Toepasbaarheid voor SIVMO. Zeer toepasbaar als voorbeeld van open data-gebaseerde populatie- en activiteitenketensynthese. De methode is inzichtelijk en herbruikbaar, maar mist de wiskundige strengheid van IPF- of generative modelling-methoden. Voor SIVMO bruikbaar als basis of benchmark in contexten met beperkte datatoegang, of als referentie voor open modelleringstrajecten.

Jain, Shubham, Nicole A. Ronald & Stephan Winter. 2015.
"*Creating a Synthetic Population: A Comparison of Tools.*" Paper presented at the 3rd Conference of Transportation Research Group of India (CTRG), December 2015. https://www.researchgate.net/publication/291608775. 587-CameraReady.pdf.

Samenvatting
Deze studie beschrijft de constructie van een synthetische populatie voor Greater Melbourne met behulp van twee vrij beschikbare softwarepakketten: PopSynWin (gebaseerd op Iterative Proportional Fitting, IPF) en PopGen (gebaseerd op Iterative Proportional Updating, IPU). Beide tools genereren microdata voor personen en huishoudens op het fijnmazige SA1-niveau (gemiddeld 400 personen), op basis van de Australische volkstelling van 2011. De populaties worden gevalideerd door de uitkomsten te vergelijken met de werkelijke aggregaten uit de census.

Het onderzoek toont aan dat beide methoden in staat zijn om populaties te genereren die sterk overeenkomen met de werkelijke verdelingen. PopGen presteert over het algemeen beter, vooral bij het tegelijkertijd benaderen van zowel persoons- als huishoudkenmerken. Verschillen worden gemeten via diverse validatiemethoden, zoals het gemiddelde verschil in distributies, kruistabellen en ruimtelijke visualisaties van afwijkingen (Delta1–Delta4).

Een belangrijke uitdaging in het proces is de voorbereiding en harmonisatie van de data, inclusief het balanceren van tegenstrijdige totalen in de aggregate datasets. Dit blijkt tijdrovend en essentieel om de synthese correct te laten verlopen. De zes gebruikte controlevariabelen (drie op huishouden- en drie op persoonsniveau) worden zorgvuldig gematcht tussen microdata en censusaggregaten.

De auteurs concluderen dat PopGen de meest geschikte tool is voor vervolggebruik in vraagmodellering, mede vanwege zijn betere prestaties op persoonsniveau. Ze pleiten voor contextspecifieke keuzes van tools en methoden, afhankelijk van de aard van de gegevens en het doel van de modellering. De synthetische populatie zal worden gekoppeld aan individuele reisdagboeken om mobiliteitsgedrag te simuleren.

Analyse

1. Methode. De auteurs vergelijken twee synthetische reconstructiemethoden: PopSynWin (IPF-gebaseerd) en PopGen (IPU-gebaseerd). Beide zijn SR-gebaseerde tools die microdata genereren voor huishoudens en personen via herhaaldelijke weging en iteratie. PopGen past huishoud- én persoonsgewichten aan in een gezamenlijke iteratieve procedure, terwijl PopSynWin alleen marginaal op huishoudniveau werkt. De methode is deterministisch met enkele afhankelijke stapsgewijze bewerkingen.

2. Inputdata. De gebruikte inputdata zijn afkomstig van de Australische volkstelling van 2011. Dit betreft:
- Microdata: 1% CURF-bestand (Confidentialised Unit Record Files)
- Aggregate data: Basic Community Profile (BCP) DataPacks op SA1-niveau

Daarnaast zijn conversietabellen gebruikt om geografische aggregatieniveaus te koppelen (SA4 → SA1). De data zijn betrouwbaar, representatief en landelijk beschikbaar.

3. Populatiekenmerken. De synthetische populatie bevat zes controlevariabelen: drie op huishoudniveau (woningtype, huishoudgrootte, aantal voertuigen) en drie op persoonsniveau (geslacht, leeftijd, arbeidsparticipatie). Dit zijn demografische en economische basiskenmerken; er zijn geen gedragskenmerken, activiteiten of verplaatsingen gegenereerd. De focus ligt op statische populatiestructuur.

4. Schaal. De populatie wordt gegenereerd voor Greater Melbourne op SA1-niveau, met een gemiddelde populatie van 400 personen per gebied. Er is geen temporele dimensie. De schaal is dus regionaal/ruimtelijk fijnmazig, maar zonder dynamiek of gedrag over tijd.

5. Output. De output bestaat uit synthetische microdata-bestanden met persoons- en huishoudkenmerken, gegenereerd op het kleinste geografische niveau. Deze data zijn bedoeld als input voor verder gebruik in modellen (zoals vraagmodellen of microsimulatie), maar het paper levert geen functionele gedragsmodellen op.

6. Validatie. Beide tools worden geëvalueerd op basis van vergelijking met werkelijke censusdistributies (verschilpercentages, kruistabellen, visualisaties). De validatie vindt plaats op verschillende ruimtelijke niveaus en voor verschillende combinaties van variabelen. PopGen levert consequent betere resultaten, vooral bij gecombineerde persoons- en huishouddistributies.

7. Openbaarheid. De gebruikte software (PopSynWin https://popsyn-win.software.informer.com/4.1/ en PopGen https://www.mobilityanalytics.org/popgen.html) is vrij beschikbaar, evenals de censusdata. De methode is reproduceerbaar mits toegang tot Australische censusdata (of een equivalente dataset) aanwezig is. De auteurs documenteren het balanceringsproces uitvoerig, inclusief beperkingen van data.

8. Toepasbaarheid voor SIVMO. De studie is methodologisch relevant voor SIVMO, vooral als basislijn of referentie voor IPF/IPU-methoden. De populatie bevat echter geen gedrag of mobiliteitsgegevens en is bedoeld als input voor verdere modellering. De nadruk ligt op betrouwbaarheid van socio-demografische representatie op klein ruimtelijk niveau. Minder geschikt voor directe gedragsimulatie, maar goed bruikbaar als basispopulatie bij gebrek aan CBS-microdata of voor scenarioverkenning op wijkniveau.

Joemanbaks, Shaya Q. J. & Jan Kiel. 2022.
The Potential Of Openstraatmap Data In Transport Models: A Case Study In Zoetermeer. Paper presented at the European Transport Conference 7-9 September 2022, Milan. ETC The Potential of OpenStreetMap Data in Transport Models - A Case Study in Zoetermeer - Joemmanbaks & Kiel.pdf

Samenvatting
In dit onderzoek is onderzocht hoe OpenStreetMap (OSM)-data gebruikt kan worden om synthetische populaties ruimtelijk te verdelen in transportmodellen, met een casestudy in Zoetermeer. Het uitgangspunt is dat gedetailleerde microdata vaak ontbreekt vanwege privacy- en beschikbaarheidsbeperkingen. Daarom wordt gebruik gemaakt van populatiesynthese via Iterative Proportional Fitting (IPF) om op basis van geaggregeerde data een gedetailleerde populatie te genereren. De koppeling van deze populatie aan daadwerkelijke woningen is een uitdaging, waarvoor OSM mogelijk een oplossing biedt.

De literatuurstudie en methodologie beschrijven de stappen om een synthetische populatie te genereren én ruimtelijk toe te wijzen aan woningen uit OSM. Deze omvatten onder meer het specificeren van het IPF-model, het harmoniseren van inputdata, de kwaliteitscontrole van OSM en het ontwikkelen van een toewijzings-methode voor huishoudens aan woningen. Voor de casestudy is OViN-data als steekproef gebruikt en CBS-data als marginaal. OSM-gegevens zijn aangevuld met veldwerk om gebouwen correct te typeren. De allocatie van huishoudens aan woningen is gedaan via regressieanalyse, gebaseerd op expertinschattingen van gewenste woonoppervlakte.

De casestudy in de wijk Meerzicht-Oost toont aan dat deze aanpak praktisch uitvoerbaar is. Hoewel externe validatie beperkt bleef, gaven vergelijkingen met WoON-data aan dat de resultaten plausibel zijn. De gebruikte methodologie bleek flexibel ondanks dat alleen opensourcedata zijn gebruikt. De toegevoegde ruimtelijke component maakt de populatie geschikt voor toepassing in microsimulatiemodellen voor vervoer.

Conclusie is dat OSM waardevol is voor het verhogen van ruimtelijke detaillering in transportmodellen. Wel is verrijking van OSM met andere databronnen (zoals BAG) aan te raden. Verdere ontwikkeling van de methode en validatie met echte microdata blijven wenselijk.

Analyse
1. Methode. De methode combineert populatiesynthese met ruimtelijke allocatie van huishoudens. Voor de populatiesynthese is Iterative Proportional Fitting (IPF) toegepast in een single-level fitting-variant, geprogrammeerd in Python. Voor de allocatie van huishoudens aan woningen is regressieanalyse gebruikt, gebaseerd op expertjudgement. OpenStreetMap (OSM) dient als databron voor de woningen.

2. Inputdata. De populatiesynthese maakt gebruik van OViN als steekproef (microdata) en CBS-data als marginaal. OSM-data is gebruikt voor het woningbestand; deze data is verrijkt via veldonderzoek. Validatiedata komt deels uit de WoON-enquête. Alle data is open source, met uitzondering van enkele bewerkte validatiesets.

3. Kenmerken. De synthetische huishoudens zijn gegenereerd op basis van drie control variables: huishoudsamenstelling, huishoudinkomen en autobezit. De woningen uit OSM zijn gefilterd en verrijkt met oppervlaktematen en aantallen wooneenheden. De koppeling van huishoudens aan woningen gebeurt via een regressiemodel waarin woonoppervlak de centrale variabele is.

4. Schaal. De casestudy beslaat één wijkniveau (Meerzicht-Oost in Zoetermeer). De schaal is dus zeer fijnmazig: een buurt binnen een Nederlandse gemeente. Dit is relevant voor modellen die disaggregatie op woningniveau vereisen.

5. Output. De uitkomst is een synthetische populatie van huishoudens met daaraan een woning toegewezen op basis van kenmerken. Deze data is geschikt als input voor microsimulatiemodellen of activity-based modellen met hoge ruimtelijke resolutie.

6. Validatie. Interne validatie is uitgevoerd via Pearson-correlatiecoëfficiënt (score: 1). Externe validatie is beperkt, deels gedaan met WoON-data. Vergelijkingen tonen kleine verschillen, maar geven geen sluitend bewijs voor accuraatheid. Validatie van de allocatiestap vond plaats via vergelijking van ruimtelijke spreiding met plausibele patronen.

7. Openbaarheid. Het onderzoek is gebaseerd op open data (CBS, OViN, OSM), en de methode is transparant beschreven. De gebruikte scripts zijn beschikbaar via de scriptie van Joemmanbaks (2022). Daarmee is de aanpak in principe reproduceerbaar.

8. Toepasbaarheid voor SIVMO. De methode is goed toepasbaar in de Nederlandse context voor stedelijke gebieden waar BAG en OSM voldoende gedetailleerd zijn. De IPF-aanpak en ruimtelijke allocatie zijn bruikbaar in agent-based modellen en microsimulaties met behoefte aan woningniveaudata. Een beperking is dat de methode nog niet is getest op landelijke schaal of in gebieden met beperkte OSM-dekking.

Kagho, Grace O., Anugrah Ilahi, Miloš Balać & Kay W. Axhausen. 2020.
"Synthetic Population of Greater Jakarta: An Iterative Proportional Updating Approach." Paper presented at the 20th Swiss Transport Research Conference (STRC), Ascona, Switzerland, May 13–15, 2020. Kagho_EtAl.pdf.

Samenvatting

Het paper beschrijft de toepassing van de Iterative Proportional Updating (IPU) methode voor het genereren van een synthetische populatie van ruim 30 miljoen personen in de regio Greater Jakarta (Jabodetabek, Indonesië). Doel is het verkrijgen van een representatieve agentpopulatie voor activity-based modellen, vooral in MATSim. De studie bouwt voort op eerder werk waarin Bayesian Networks en Generalized Raking (multilevel IPF) zijn gebruikt, maar verbetert die aanpak door zowel personen- als huishoudkenmerken gelijktijdig te matchen met marges op het laagste geografische niveau (subdistricts).

De IPU-methode, ontwikkeld door Ye et al. (2009), lost bekende beperkingen van IPF op. IPU maakt het mogelijk om de verdeling van personen- en huishoudkenmerken onafhankelijk aan te passen, zonder inconsistenties in weging. In de studie worden censusdata op subdistrictniveau (n=1.336) gecombineerd met een grote household travel survey (JAPTRAPIS, n≈657.000 individuen, n≈179.000 huishoudens). Marges uit de census (2016–2018) zijn geharmoniseerd, en ontbrekende of foutieve waarden zijn aangepast met behulp van regioverdelingen, historische data of plausibiliteitsregels.

PopGen 2.0 (Python CLI) is gebruikt voor de implementatie. De simulatie duurde 27 uur voor 1.000 iteraties. Validatie toont een bijna perfecte overeenkomst met censuscijfers: $R^2 = 0.999$ voor zowel personen als huishoudens. Op subdistrictniveau zijn de relatieve afwijkingen beperkt: minder dan 5% van de zones heeft een fout >1% in totaal aantal personen. Alleen de leeftijdsgroep 65+ kent grotere afwijkingen (tot 18,8%), vermoedelijk door ondervertegenwoordiging in de steekproef.

De studie laat zien dat IPU goed toepasbaar is voor grote, gedetailleerde regio's met beperkte inputkwaliteit. Er wordt gepleit voor uitbreiding met meer huishoud-kenmerken (zoals autobezit, inkomen) en voor kwaliteitscontrole op de inputdata. De gegenereerde populatie is bruikbaar als input voor MATSim.

Analyse
1. Methode. De studie maakt gebruik van de *Iterative Proportional Updating* (IPU) methode. Deze methode is een uitbreiding van IPF en stelt gebruikers in staat om zowel huishoud- als persoonskenmerken gelijktijdig te matchen aan marginale totalen uit de censusdata. IPU is robuuster voor kleinere geografische eenheden en voorkomt het oneigenlijk overnemen van huishoudgewichten voor persoonskenmerken. De implementatie is uitgevoerd met PopGen 2.0, een Python-gebaseerd open source tool.

2. Inputdata. Twee databronnen worden gecombineerd: een grootschalige huishoudverplaatsingsenquête (JAPTRAPIS, 2012) met ca. 179.000 huishoudens en 657.000 individuen, en geaggregeerde censusdata (2016–2018) voor subdistricten in Groot-Jakarta. Voor ontbrekende of inconsistente data worden pragmatische

oplossingen toegepast, zoals toewijzing op basis van hogere geografische niveaus en aangepaste groepering van leeftijdsklassen.

3. Kenmerken. Zowel huishoudkenmerken (zoals huishoudgrootte) als persoons-kenmerken (leeftijd en geslacht) worden gemodelleerd. Verdere attributen uit de enquête (zoals inkomen, voertuigbezit en werkstatus) zijn via imputatie aan de synthetische populatie toegevoegd, maar zijn niet gebruikt als controlevariabelen in het syntheseproces.

4. Schaal. De synthetische populatie bestrijkt het volledige stedelijke gebied van Groot-Jakarta (~30 miljoen inwoners) en is gegenereerd op het niveau van 1.336 subdistricten (gemiddeld 10.000 inwoners per zone). Dit is uitzonderlijk gedetailleerd voor een studie in een megastad.

5. Output. Het resultaat is een volledige synthetische populatie met matching op persoons- en huishoudniveau. De matching toont een bijna perfecte fit ($R^2 = 0.999$). De afwijkingen blijven in de meeste subdistricten onder 1–3%, met grotere afwijkingen (>10%) slechts voor specifieke leeftijdsgroepen (vooral 65+).

6. Validatie. Validatie is uitgevoerd op meerdere niveaus: totalen, verdelingen per attribuut, en ruimtelijke spreiding van fouten. Er zijn histogrammen en boxplots gebruikt om de afwijkingen per subdistrict en leeftijdsgroep te tonen. De resultaten zijn transparant gepresenteerd, inclusief foutmarges per attribuut.

7. Openbaarheid. Het gebruikte IPU-algoritme is geïmplementeerd met het open-source programma PopGen 2.0. De JAPTRAPIS-enquête is niet openbaar, maar werd beschikbaar gesteld door JICA. De censusdata komt uit publieke overheidsbronnen, hoewel in de praktijk moeilijk toegankelijk.

8. Toepasbaarheid voor SIVMO. Deze aanpak is goed overdraagbaar naar andere contexten, mits een redelijke steekproef en voldoende gedetailleerde censusdata beschikbaar zijn. De toepassing op subdistricten sluit aan bij de wens binnen SIVMO om op laag schaalniveau betrouwbare populaties te genereren. PopGen 2.0 is inzetbaar, al vereist het handmatige datavoorbewerking en kennis van Python. De beperking zit vooral in het aantal controlevariabelen; uitbreiding vereist aanvullende data en meer verwerkingscapaciteit.

Kang, Jaewoong, Young Kim, Muhammad Mu'az Imran, Gi-sun Jung & Yun Bae Kim. 2023.
"*Generating Population Synthesis Using a Diffusion Model.*" In Proceedings of the 2023 Winter Simulation Conference, edited by C.G.. Corlu et al., 2944–2955. IEEE. 2023-kang-kim-imran-jung-kim.pdf.

Samenvatting

In deze studie wordt een nieuwe methode voorgesteld om synthetische populaties te genereren met behulp van een *denoising diffusion probabilistic model* (DDPM), een type deep generatief model. Traditionele methoden zoals IPF en MCMC kampen met schaalproblemen en het niet kunnen genereren van zeldzame of ontbrekende combinaties van kenmerken (sampling zeros). DDPM daarentegen bouwt voort op technieken uit beeldsynthese, waarbij gegevens eerst worden omgezet in ruis en vervolgens stapsgewijs terugvertaald naar plausibele populatiegegevens. Deze methode biedt potentie om sampling zeros te genereren zonder de structurele consistentie van de populatie aan te tasten.

Het artikel vergelijkt DDPM met bestaande technieken zoals VAE en MCMC op basis van synthetische populaties voor Korea. De resultaten tonen aan dat DDPM lagere foutmarges heeft (SRMSE = 2.13) dan VAE (2.38) en MCMC (7.62). Ook laat DDPM een betere dekking zien van sampling zeros, terwijl het aantal structurele fouten beperkt blijft dankzij rule-based post-processing met censusregels. De aanpak vereist echter veel rekenkracht en handmatige behandeling van structurele restricties via bins afgeleid uit censusdata.

De methode werd getest met 2% steekproefdata van het Koreaans bureau voor statistiek, waarbij zowel persoons- als mobiliteitskenmerken zijn meegenomen. De data zijn geprepareerd en getransformeerd tot vierkante matrices om compatibel te zijn met het DDPM-trainingsproces. Training gebeurde op een GPU met 1000 stappen en 300 epochs. Na training werden synthetische populaties gegenereerd en vergeleken met echte populaties via meerdere foutmaten.

De auteurs concluderen dat DDPM een veelbelovende aanpak is voor grootschalige populatiesynthese, met name voor het oplossen van sampling zeros en schaalproblemen. Ze wijzen erop dat toekomstige uitbreiding nodig is richting huishoudkenmerken en efficiëntere verwerking van structurele restricties. De studie toont aan dat innovatieve deep learning-methoden uit andere domeinen, zoals beeldherkenning, effectief kunnen worden ingezet voor sociaal-wetenschappelijke toepassingen zoals populatiesynthese.

Analyse
1. Methode. De auteurs gebruiken een *denoising diffusion probabilistic model* (DDPM), een geavanceerd deep generatief model dat oorspronkelijk is ontwikkeld voor beeldgeneratie. De methode bouwt voort op het iteratief verstoren en reconstrueren van gegevens met behulp van een getraind neuraal netwerk. Dit model genereert synthetische populatiegegevens vanuit ruis, waarmee het sampling zero-probleem effectief wordt aangepakt. Structurele fouten (onmogelijke combinaties) worden verwijderd via rule-based postprocessing, op basis van bekende censusregels.

2. Inputdata. De studie gebruikt een 2%-steekproef van de Koreaanse microdata (927.843 individuen), afkomstig van KOSTAT. Daarnaast worden censusgegevens gebruikt voor het definiëren van realistische marges en het verwijderen van structurele fouten. De input bevat demografische kenmerken (leeftijd, geslacht), woonplaatscodes, mobiliteitskenmerken en vervoermiddelen. Household data is nog niet meegenomen, wat door de auteurs als beperking wordt erkend.

3. Kenmerken. De gegenereerde populatie omvat individuele kenmerken zoals leeftijd, geslacht, woonlocatie (provincie en stad), reistijd, vervoermiddel en woon-werkrelaties. De focus ligt dus op persoonsniveau. Structurele beperkingen op numerieke waarden (zoals leeftijd ≥ 0) zijn handmatig ingebouwd. Huishoud-kenmerken ontbreken, maar uitbreiding is voorzien in vervolgonderzoek.

4. Schaal. De schaal is nationaal (Zuid-Korea), waarbij gebruikgemaakt wordt van 2% microdata. De methode is in principe schaalbaar naar grotere datasets en andere landen, mits er voldoende representatieve steekproeven en censusdata beschikbaar zijn. Trainingstijd en rekencapaciteit vormen hierbij een praktische beperking.

5. Output. De output bestaat uit synthetische individuen met plausibele combinaties van kenmerken, gegenereerd via DDPM. De kwaliteit van de output wordt gemeten met SRMSE, waarin de DDPM beter scoort dan MCMC en VAE. Sampling zeros worden succesvol gegenereerd; structurele zeros worden grotendeels vermeden via bins.

6. Validatie. Validatie gebeurt met de *standardized root mean squared error* (SRMSE), op zowel marginale als bivariate verdelingen. De DDPM scoort gemiddeld 2.13, beter dan VAE (2.38) en MCMC (7.62). De afstand tot echte populatiepunten wordt ook gemeten. Er is visuele inspectie (grafieken), maar geen validatie op extern gedrag of reispatronen.

7. Openbaarheid. De code is publiek beschikbaar via GitHub (https://github.com/Ninekrad/PopulationSynthesis). De gebruikte data (2% microdata van KOSTAT) is in principe beperkt toegankelijk vanwege privacy, maar publiek voor onderzoeksdoeleinden in Korea. De gebruikte censusregels voor postprocessing zijn gebaseerd op publieke statistieken.

8. Toepasbaarheid voor SIVMO. De methode is vernieuwend en relevant voor SIVMO in termen van sampling zeros, hoge dimensionering en deep learning. Echter, de vereiste rekenkracht, het ontbreken van huishoudstructuren en de huidige afstemming op Koreaanse data beperken directe inzetbaarheid. De aanpak is vooral interessant als experimentele benchmark of voor gespecialiseerde toepassingen binnen grote ABM's, mits voldoende (open) data beschikbaar is. Integratie met bestaande populatiegeneratoren in Nederland zou extra ontwikkelwerk vragen.

Kouwenhoven, Marco & Dylan Mulders. 2022.
Opzet populatiesimulator. Memo 20025-M27 v6, June 7, 2022. The Hague: Significance. M27_-_opzet_populatiesimulator_-_v9.pdf.

Deze memo beschrijft de technische werking van een populatiesimulator die op persoons- en huishoudniveau de demografische en sociaal-economische ontwikkeling van Nederland simuleert. De simulator is opgebouwd uit negen hoofdstappen en vier deelstappen, die jaarlijks worden doorlopen: initialisatie, overlijden, geboorte, verandering huishoudsamenstelling, toevoegen huishoudens, verhuizing, verandering van sociaal-economische status, verandering van sector en verandering van inkomen.

Elke stap is gebaseerd op overgangskansen (per individu of huishouden) en op targets per jaar, regio en klasse. Voor overlijden en geboorte zijn de overgangskansen afgeleid uit Pearl en Tigris (WLO-Laag en -Hoog gemiddeld). Targets worden per scenario opgegeven en toegepast als harde sturing: als er bijvoorbeeld 100.000 geboortes moeten plaatsvinden, worden precies zoveel geboorteacties toegekend aan de hoogste kansen in de populatie.

Bij huishoudtransities worden vier processen onderscheiden: het vormen van eenpersoonshuishoudens, samenwonen na vertrek uit ouderlijk huis, samenwonen binnen het huishouden, en scheidingen. Voor elk proces gelden specifieke regels, gebaseerd op leeftijd, geslacht en andere kenmerken. Er wordt expliciet gelet op evenwichten tussen partners, om bijvoorbeeld over- of onderproductie van huishoudtypes te voorkomen.

Om migratie te modelleren, worden huishoudens toegevoegd of verwijderd, op basis van CBS-cijfers per leeftijdsgroep. Andere transities (status, sector, inkomen) zijn gebaseerd op rekenregels of overgangstabellen, deels afgeleid van CBS-analyse en LMS-SEGS. Voor sociaal-economische status is een aparte submodule uitgewerkt met meerdere afhankelijkheden (zoals leeftijd, vorige status, pensioenleeftijd). Het resultaat is een jaarlijks geüpdatet populatiebestand met consistente kenmerken.

Belangrijke eigenschap van het systeem is dat het niet werkt met klassieke stochastische trekking per individu, maar met een variantie-reductietechniek: rangordes op basis van kans, gevolgd door selectie totdat de doelwaarde (target) bereikt is. Dit vergroot reproduceerbaarheid en maakt het model deterministisch gegeven gelijke input.

De simulator is ontworpen voor toepassing in het SPARK-model, maar is gebaseerd op bestaande componenten uit Pearl en Tigris. De structuur is modulair, herhaalbaar en in hoge mate data-gedreven.

Analyse
1. Methode. De populatiesimulator is een regelgebaseerd microsimulatiemodel dat jaarlijks de status van personen en huishoudens bijwerkt via overgangskansen. De modelstappen zijn gedetailleerd uitgewerkt per type gebeurtenis (geboorte, overlijden, verhuizing, etc.) en maken gebruik van variantie-reductie via probabilistische trekking binnen scenario-gebonden targets. De aanpak is deterministisch op het niveau van de uitkomst (via targets), maar stochastisch op het niveau van toewijzing. Er is sprake van iteratieve verwerking over individuen of huishoudens, waarbij transities plaatsvinden onder expliciete kansstructuren.

2. Inputdata. Het model gebruikt demografische microdata van personen en huishoudens als input, inclusief kenmerken zoals leeftijd, geslacht, huishoudpositie, regio en stedelijkheidsgraad. Externe databronnen zijn onder meer: CBS-prognoses (bevolking, sterfte, geboortes, arbeidsmarkt, inkomens), Pearl/Tigris overgangskansen voor levensloopgebeurtenissen, LMS-SEGS voor sociaaleconomische status, sector en verhuiskansen, CBS-microdata voor sector- en verhuisanalyses. Scenario's zoals WLO-Laag en WLO-Hoog bepalen de targets voor aantallen en verdelingen.

3. Kenmerken. De simulator bevat een groot aantal persoons- en huishouddimensies: leeftijd, geslacht, partnerstatus, kindertal, regio, stedelijkheid, sociaal-economische status, sector, inkomen, enzovoorts. De huishoudstructuur (aantal hoofdpersonen, kinderen, positie) is integraal onderdeel van het model. De structuur is zodanig opgezet dat consistentie tussen persoons- en huishouddimensies behouden blijft.

4. Schaal. De simulator is ontworpen voor toepassingen in Nederland, met eenheden op persoons- en huishoudniveau en aggregatie naar regio's en stedelijkheidsklassen. Schaalbaarheid lijkt technisch mogelijk, maar hangt af van beschikbaarheid van microdata en doelscenario's. De toepassing is vooral geschikt voor strategische langetermijnscenario's (tot 2060).

5. Output. De output bestaat uit jaarlijkse gesynthetiseerde microdata-bestanden met voor elke persoon en elk huishouden een set geactualiseerde kenmerken. Hiermee kunnen aggregaties worden gemaakt naar populatiekenmerken per jaar en per regio, en ook beleidsspecifieke indicatoren zoals arbeidsmarktparticipatie, huishoudgroei en inkomensverdeling. Er is controle op overeenstemming met externe targets.

6. Validatie. Er is geen expliciete beschrijving van formele validatieprocedures, maar de consistentie met CBS-prognoses en LMS-SEGS is ingebouwd via doelgerichte kalibratie op externe targets. Het model past een variantie-reductietechniek toe om de overgangskansen zo toe te passen dat outputaantallen in lijn zijn met de gewenste scenarioresultaten. Validatie vindt impliciet plaats via matching aan bekende totalen.

7. Openbaarheid. De inhoud van het model, inclusief methodiek en gebruikte gegevens, is goed gedocumenteerd. De implementatie (code) is echter niet openbaar, en de gebruikte data (bijvoorbeeld microdata van het CBS) zijn niet vrij toegankelijk. Daardoor is volledige reproduceerbaarheid voor externe partijen beperkt.

8. Toepasbaarheid voor SIVMO. De simulator is goed toepasbaar voor beleidsmatige en demografisch-sociaal georiënteerde toepassingen in het kader van mobiliteitsmodellen, mits voldoende aansluiting is met verkeerskundige modellen. Vooral de koppeling met huishoudstructuren, regio's en inkomensniveaus is relevant. Het model is geschikt voor het genereren van populaties voor SIVMO-achtige modellen, mits uitbreiding of koppeling met mobiliteitskenmerken wordt voorzien.

Kukic, Marija & Michel Bierlaire. 2021.
"*The Case of Population Synthesis at the Level of the Households*." Paper presented at the 21st Swiss Transport Research Conference (STRC), Ascona, September 12–14, 2021. Kukic_Bierlaire.pdf.

Samenvatting

Dit paper beschrijft de ontwikkeling van een methode voor het synthetisch genereren van huishoudens waarbij personen en huishoudens gelijktijdig worden gegenereerd. De auteurs signaleren dat bestaande populatiesynthesemethoden vaak in twee stappen werken: eerst worden individuen gegenereerd, vervolgens worden zij in huishoudens geplaatst. Dit leidt tot inconsistenties en onrealistische combinaties, zoals kinderen die ouder zijn dan hun ouders.

De voorgestelde methode is gebaseerd op een Gibbs sampler (een Markov Chain Monte Carlo-techniek) waarbij conditionele verdelingen gebruikt worden om kenmerken van individuen binnen een huishouden te genereren, inclusief beperkingen op basis van domeinkennis. Zo kan bijvoorbeeld worden gewaarborgd dat kinderen jonger zijn dan hun ouders, of dat inkomens en werkstatus overeenkomen met leeftijd en opleiding.

De methode wordt toegepast als 'imputatiecomponent' binnen een groter onderzoeksproject (Multi-day and Multi-person Activity Patterns and Schedules Owners). De populatie wordt gegenereerd door gegevens uit twee datasets te combineren: de MOBIS-studie (mobiele app-data, 2019) en de Zwitserse mobiliteitsenquête (censusdata, 2015). Vanuit de MOBIS-dataset worden 'anker-individuen' gekozen en vervolgens worden huishoudens gegenereerd op basis van censusdistributies.

Het paper beschrijft gedetailleerd hoe huishoudens met verschillende structuren worden gegenereerd (single, paar zonder kinderen, paar met kinderen, eenoudergezinnen, niet-familiaire huishoudens). Validatie vindt plaats met behulp van SRMSE en R^2 op basis van age- en genderverdelingen. Ook wordt vergeleken met TGANs (tabular generative adversarial networks), waarbij wordt geconcludeerd dat de Gibbs-methode realistischer data oplevert ondanks iets minder nauwkeurige marges.

Analyse

1. Methode. De auteurs ontwikkelen een simulatiegebaseerde methode voor populatiesynthese op huishoudenniveau, waarbij generatie van individuen en hun toewijzing aan huishoudens in één stap gebeurt. De kernmethode is gebaseerd op *Gibbs sampling* binnen een Markov Chain Monte Carlo (MCMC) framework. De benadering integreert domeinkennis door conditionele verdelingen te construeren waarin realistische beperkingen zijn ingebouwd (bijv. leeftijdsverhoudingen tussen ouders en kinderen).

*2. Inputdata.*Twee datasets worden gebruikt: (1) de Zwitserse mobiliteitsenquête (MTMC 2015) als referentie voor conditionele verdelingen en (2) de MOBIS 2019 dataset als doelbestand waarin synthetische huishoudens worden geïmputeerd. De datasets bevatten sociaaleconomische kenmerken op persoons- en huishouden-

niveau, waaronder leeftijd, geslacht, opleiding, inkomen, huishoudgrootte, huishoudtype en autobezit.

3. Kenmerken. De methode genereert populaties waarin persoons- en huishoudenkenmerken consistent en realistisch gecombineerd zijn. Door regels op te nemen in de constructie van conditionele verdelingen (zoals 'geen kind ouder dan de ouder'), wordt gezorgd voor plausibele resultaten. De procedure is deels stochastisch (getrokken uit verdelingen) en deels deterministisch (kenmerken overgenomen uit bestaande rijen).

4. Schaal. De aanpak is toegepast op Zwitserland als geheel, met synthetische uitbreiding van de MOBIS dataset van 3.700 naar 10.736 personen. De methode is schaalbaar, zolang voldoende referentiegegevens beschikbaar zijn, en zou toepasbaar zijn op regionale of nationale schaal.

5. Output. De gegenereerde populatie bevat huishoudens met meerdere leden, waarbij per persoon elf attributen worden gespecificeerd. De auteurs rapporteren nauwkeurige overeenkomsten met de referentiegegevens op marginale verdelingen, maar onderstrepen vooral de interne consistentie van huishoudstructuren. Validatiestatistieken (SRMSE, R^2) worden gebruikt om de kwaliteit te beoordelen.

6. Validatie. De validatie is zowel kwantitatief (marginale distributies, SRMSE, R^2) als kwalitatief (controle op onrealistische combinaties zoals kinderen met inkomen). De auteurs tonen dat hun methode robuuster is dan GAN-gebaseerde technieken, die weliswaar betere marges genereren, maar vaker onrealistische rijen produceren.

7. Openbaarheid. De paper beschrijft de methode en bevat pseudocode, maar geen volledige broncode of downloadbare tool. De onderliggende datasets zijn (semi-) publiek, maar bewerking en koppeling zijn handmatig uitgevoerd. Daarmee is de aanpak reproduceerbaar in principe, maar vergt replicatie substantiële inspanning.

8. Toepasbaarheid voor SIVMO. De aanpak is relevant voor SIVMO vanwege de integratie van huishoudstructuren, de controleerbaarheid van de synthetische gegevens en de focus op realisme. De methode biedt een bruikbare aanvulling op bestaande technieken zoals IPF/IPU, vooral voor toepassingen waarbij huishoud-interacties een rol spelen (zoals activity- en agent-based modellen). Nadelen zijn de afhankelijkheid van gedetailleerde referentiegegevens en het ontbreken van een gebruiksklare software-implementatie.

La, Duc Minh & Hai L. Vu. 2024.
"*A Pool-Based Approach to Population Synthesis in Transport Modeling."* Paper to be presented at the Australasian Transport Research Forum 2024, Melbourne, Australia, November 27–29, 2024. ATRF2024_Abridged_51-1.pdf.

Samenvatting

La en Vu stellen in dit paper een nieuwe methode voor populatiesynthese voor: Sequential Attribute Adjustment (SAA), gebaseerd op een zogeheten *pool-based approach*. Deze methode beoogt de tekortkomingen van traditionele technieken zoals Iterative Proportional Fitting (IPF) en Bayesian Networks (BN) te verhelpen. Klassieke technieken hebben moeite met de 'curse of dimensionality', het 'zero-cell'-probleem en het bewaren van realistische combinaties van kenmerken.

De kern van SAA is dat een initiële pool van synthetische agents wordt gegenereerd uit surveydata (seed data), bijvoorbeeld via een generatief model zoals een Bayesian Network. Vervolgens worden attributen in deze pool sequentieel aangepast om te voldoen aan de marginale verdelingen uit censusdata, waarbij de onderlinge afhankelijkheden tussen kenmerken behouden blijven. Dit verschilt van IPF, dat alleen werkt met marginale gewichten, en van BN, dat vaak onrealistische combinaties oplevert.

De methode is getest op huishouddata voor de staat Victoria (Australië), waarbij vijf attributen worden gesynthetiseerd: huishoudgrootte, aantal voertuigen, huishoudinkomen, woningtype en eigendomstatus. Aggregaten zijn afkomstig uit de Australische volkstelling (2021), terwijl de disaggregaten uit VISTA (2012–2018) komen. De evaluatie gebruikt RMSE en Jensen-Shannon Divergence (JSD) om nauwkeurigheid en distributiebehoud te meten. SAA presteert beter dan IPF en BN, vooral bij attributen met zeldzame of ontbrekende combinaties. In het geval van voertuigaantal bijvoorbeeld behaalt SAA een RMSE van 2,31 tegenover 8,81 (IPF) en 208,46 (BN).

De methode lost ook effectief het zero-cell-probleem op, bijvoorbeeld door realistische combinaties te genereren voor huishoudens met negatieve inkomens die ontbreken in de originele steekproef. De auteurs bepleiten verdere uitbreiding naar andere attributen en grootschalige toepassingen.

Analyse

1. Methode. De paper introduceert *Sequential Attribute Adjustment (SAA)* als nieuwe methode voor populatiesynthese. De kern is een *pool-based* benadering waarbij synthetische individuen uit seed data worden gevormd en hun attributen sequentieel worden aangepast om te voldoen aan marginale distributies. Elke stap houdt rekening met eerder aangepaste attributen, waardoor onderlinge afhankelijkheden behouden blijven. De initiële pool wordt gegenereerd met een generatief model, zoals een Bayesian Network, waarna stapsgewijze correctie plaatsvindt.

2. Inputdata. SAA gebruikt zowel geaggregeerde data uit de Australische volkstelling van 2021 (census) als gedesaggregeerde surveydata uit VISTA (2012–2018). De toepassing richt zich op huishoudkenmerken: grootte, aantal voertuigen, inkomen, woningtype en eigendomssituatie. De combinatie van datasets stelt hoge eisen aan reconciliatie van inconsistenties, iets wat SAA expliciet adresseert.

3. Populatiekenmerken. De synthetische populatie bevat vijf huishoudkenmerken op postcodeniveau (691 POA's). De methode is flexibel genoeg om complexere of ontbrekende combinaties te genereren (bijv. huishoudens met negatieve inkomens). Doordat het algoritme sequentieel werkt en rekening houdt met onderlinge afhankelijkheden, wordt realisme in de populatiestructuur bevorderd.

4. Schaal. De methode is toegepast op deelstaatniveau (Victoria, Australië), gedetailleerd tot postcodegebieden. Dit toont geschiktheid voor grootschalige toepassingen. De auteurs claimen dat de methode schaalbaar is naar grotere gebieden of extra kenmerken, wat belangrijk is voor landelijke of regionale toepassingen in Nederland.

5. Output. De output is een gesynthetiseerde huishoudpopulatie waarin marges nauwkeurig overeenkomen met censusdata, terwijl de verdelingen van de oorspronkelijke seed data grotendeels behouden blijven. De methode voorkomt bovendien het *zero-cell* probleem, wat leidt tot realistische combinaties ook bij zeldzame categorieën. Resultaten worden geëvalueerd met RMSE en Jensen-Shannon Divergence.

6. Validatie. SAA is vergeleken met IPF en BN aan de hand van RMSE en JSD. In alle gevallen presteerde SAA beter, bijvoorbeeld met een RMSE van 2,31 voor voertuigen (vs. 8,81 voor IPF en 208,46 voor BN). Ook in het behoud van oorspronkelijke verdelingen (lage JSD) scoort SAA goed. Validatie is systematisch opgezet en kwantitatief onderbouwd.

7. Openbaarheid. De paper beschrijft de methode duidelijk, maar biedt geen directe toegang tot broncode of softwaretool. De gebruikte datasets zijn in principe openbaar (ABS census, VISTA), al is het verwerken van die data niet triviaal. De methode lijkt reproduceerbaar mits kennis van de techniek en toegang tot beide databronnen.

8. Toepasbaarheid voor SIVMO. SAA is relevant voor SIVMO vanwege het vermogen om marges en verdelingen te combineren, realistische populaties te genereren én zeldzame categorieën te modelleren. De aanpak biedt controle, transparantie en aanpasbaarheid. SAA lijkt vooral geschikt voor toepassingen waarbij meerdere databronnen gecombineerd moeten worden of waar traditionele methoden zoals IPF tekortschieten.

Lim, Poh Ping & David Gargett. 2013.
"Population Synthesis for Travel Demand Forecasting." In Proceedings of the 36th Australasian Transport Research Forum (ATRF), October 2–4, 2013, Brisbane, Australia. 2013_lim_gargett.pdf.

Samenvatting

In dit paper presenteren Lim en Gargett een toepassing van populationsynthese ten behoeve van activity-based microsimulatiemodellen voor het voorspellen van vervoervraag in Australische steden. De studie beschrijft de constructie van een synthetische populatie voor de Greater Metropolitan Area van Sydney op het niveau van Travel Zones (TZ) en Census Collection Districts (CCD) op basis van de volkstelling van 2006. De belangrijkste motivatie is het ontbreken van publiek beschikbare microdata op persoons- en huishoudenniveau, wat met behulp van synthetische data kan worden aangevuld.

De auteurs gebruiken de Iterative Proportional Updating (IPU) methode, een uitbreiding van de traditionele Iterative Proportional Fitting (IPF), waarmee simultaan zowel huishouden- als persoonsattributen worden gesynthetiseerd. Dit gebeurt op basis van 1%-steekproefdata (CURFs) en geaggregeerde censusdata van het Australisch Bureau voor Statistiek. Om consistentie tussen datasets en geografische indelingen te waarborgen, wordt een uitgebreide 'balancing'-procedure uitgevoerd.

PopGen, een standalone synthesizer ontwikkeld aan Arizona State University, wordt ingezet voor het genereren van de populatie. Dit vereist een aanpassing van Australische geografische indelingen aan de Amerikaanse PUMA-structuur. De validatie van de synthetische populatie laat op TZ- en CCD-niveau goede overeenkomsten zien met de echte censusdata: verschillen in aantallen personen liggen tussen 1,9% en 3,45%.

Het paper bespreekt ook bekende knelpunten, zoals het 'zero cell'-probleem, en de noodzaak voor verdere validatie, vooral op persoonsniveau. In de vervolgfasen van het project zal de synthetische populatie gecombineerd worden met enquêtegegevens over verplaatsingsgedrag (HTS) om reële activiteits- en verplaatsingspatronen te simuleren.

Analyse
1. Methode. De studie past de *Iterative Proportional Updating* (IPU)-methode toe, een uitbreiding van de klassieke IPF-techniek. IPU maakt het mogelijk om zowel huishoud- als persoonskenmerken gelijktijdig af te stemmen op marginale verdelingen uit censusdata. De procedure is geïmplementeerd via PopGen-software, die iteratief gewichten toekent op basis van een steekproef (1% CURF) en marges uit de census. Er wordt aandacht besteed aan balancering van gegevens, geografische herindeling en hercategorisering van variabelen.

2. Inputdata. Twee databronnen zijn gebruikt: (1) de 1% CURF uit de Australische volkstelling 2006 (microdata) en (2) geaggregeerde marges uit de Census Table Builder van het ABS. Het studiegebied is Sydney GMA, met 1.953 Travel Zones (TZ) en 8.374 Census Collection Districts (CCD). Voor controle zijn variabelen als huishoudenstype, woningtype, autobezit, leeftijd, geslacht en arbeidsmarktpositie geselecteerd.

3. Kenmerken. De gesynthetiseerde populatie bevat zowel huishoud- als persoonskenmerken. Voor huishoudens omvat dit onder andere grootte, structuur en voertuigbezit; voor personen gaat het om leeftijd, geslacht en arbeidsmarktstatus. De populatie is gesimuleerd op het niveau van TZ's en CCD's. Er is expliciete aandacht voor het behoud van correlaties uit de seed data.

4. Schaal. De methode is toegepast op een grootstedelijk gebied met hoge ruimtelijke resolutie (CCD-niveau). Daarmee is het model inzetbaar voor kleinschalige gebiedsdekkende analyses, iets wat relevant is voor lokale vervoervraagvoorspelling. Er wordt ook verwezen naar bredere toepassing voor Australische hoofdsteden.

5. Output. De output bestaat uit vier gesynthetiseerde datasets (huishoudens en personen op CCD- en TZ-niveau). Validatie laat kleine afwijkingen zien: op CCD-niveau is het verschil in aantal personen 1,9%, op TZ-niveau 3,45%. Ook op kenmerkdistributie is de overeenkomst met censusdata overwegend goed (verschillen <1% voor meeste categorieën).

6. Validatie. Validatie is kwantitatief uitgevoerd via vergelijking van marges (percentuele afwijkingen) en gemiddelde absolute afwijkingen per zone (d-waarde). De spreiding van deze afwijkingen is in figuren weergegeven, en resultaten per kenmerk zijn nauwkeurig in tabellen opgenomen. Er wordt gewezen op enkele resterende afwijkingen, vooral bij persoonskenmerken op TZ-niveau.

7. Openbaarheid. De gebruikte datasets zijn deels openbaar (Census Table Builder), deels beperkt beschikbaar (CURF). De gebruikte software (PopGen, https://www.mobilityanalytics.org/popgen.html) is openbaar, maar ontwikkeld voor de VS; aanpassingen waren nodig voor toepassing op Australische gegevens. De beschrijving van het IPU-proces is reproduceerbaar, maar de gebruikte scripts zijn niet gedeeld.

8. Toepasbaarheid voor SIVMO. De methode is toepasbaar voor SIVMO, vooral vanwege het gebruik van IPU en de verfijnde toepassing op klein-schalig niveau. De aanpak is goed overdraagbaar op Nederlandse situaties met CBS-marges en microdata. Wel is er aandacht nodig voor het *zero-cell*-probleem en de afstemming tussen microdata en marges. De praktische uitvoerbaarheid (balancering, herstructurering) en schaalbaarheid zijn positief.

Müller, Kirill. 2014.
A Generalized Approach to Population Synthesis. PhD diss., ETH Zürich.
https://doi.org/10.3929/ethz-b-000171586. phd-thesis-muelleki-final.pdf.

Het proefschrift *A Generalized Approach to Population Synthesis* van Kirill Müller richt zich op methoden voor het genereren van synthetische huishoudpopulaties ten behoeve van agent-based microsimulaties in transportplanning. Hoewel veel onderzoek is gedaan naar de synthese van individuele personen, is het modelleren van hele huishoudens minder ver ontwikkeld. Dit werk behandelt drie methoden—Iterative Proportional Updating (IPU), Entropy Optimization (EO) en de nieuw voorgestelde Hierarchical Iterative Proportional Fitting (HIPF)—en plaatst deze in één algoritmisch kader. Daarbij wordt aangetoond dat EO wiskundig equivalent is aan Generalized Raking, wat leidt tot efficiëntere en beter onderbouwde toepassingen.

De kernbijdrage van het proefschrift is een herformulering van het synthetiseren van populaties als een optimalisatieprobleem over continue vectoren. Dit maakt het mogelijk om bestaande methoden te vereenvoudigen en te verbeteren, en opent de deur naar het gebruik van statistische technieken zoals 'generalized raking'. Door deze aanpak wordt ook een snelle toets op de haalbaarheid van de opgelegde randvoorwaarden mogelijk. Voor de implementatie wordt gebruik gemaakt van open-source software in de R-omgeving, waarmee het hele proces reproduceerbaar en transparant wordt.

De methoden zijn toegepast op de bevolkingsgegevens van Zwitserland. Hieruit blijkt dat generalized raking superieur presteert ten opzichte van bestaande methoden: het vereist minder rekentijd en geeft betere aansluiting op de invoergegevens. De methode wordt ook succesvol gebruikt om activiteitsketens aan te passen aan toekomst scenario's, bijvoorbeeld voor scenario's in het MATSim-vervoersmodel.

Tot slot bepleit Müller dat toekomstige onderzoek zich zou moeten richten op geschikte datatransformaties, waarmee standaardstatistische methoden eenvoudiger inzetbaar worden, in plaats van telkens nieuwe algoritmes te ontwikkelen.

Analyse
1. Methode. De studie introduceert en vergelijkt drie methoden voor populatiesynthese: Iterative Proportional Updating (IPU), Hierarchical Iterative Proportional Fitting (HIPF) en Entropy Optimization (EO). Een belangrijke bijdrage is de herformulering van EO als generalized raking, waarmee een statistisch stevig gefundeerde aanpak ontstaat. Deze methoden worden gepresenteerd binnen een uniform algoritmisch kader. Er is nadruk op theoretische onderbouwing, herformuleerbaarheid in vectorvorm en het gebruik van optimalisatietechnieken.

2. Inputdata. De methoden maken gebruik van steekproeven van huishoudens (bij voorkeur met volledige huishoudstructuur), in combinatie met controle-tellingen op zowel huishoud- als persoonsniveau. Voor empirische toepassing is gebruikgemaakt van de Zwitserse bevolkingscensus, inclusief gedetailleerde kenmerken en activiteitsketens.

3. Kenmerken. Er worden zowel huishoudkenmerken als persoonskenmerken gemodelleerd, inclusief hiërarchische structuur (bijv. meerdere personen binnen één huishouden). Ook gedragsdata (activiteitsketens) worden betrokken in het

synthetiseren van een plausibele populatie. Gewogen steekproeftrekking zonder terugleggen speelt hierbij een rol.

4. Schaal. De methoden zijn generiek toepasbaar, maar zijn gevalideerd op landelijke schaal (Zwitserland). Er is geen beperking in geografische resolutie zolang voldoende controledata beschikbaar zijn. De benadering is schaalbaar naar grotere datasets door gebruik van efficiënte algoritmen.

5. Output. De output is een synthetische populatie van huishoudens en personen die statistisch overeenkomt met de ingegeven marges. Daarbij worden zowel attributen als gedragskenmerken (activiteitspatronen) gegenereerd. Er is aandacht voor realistische interne consistentie binnen huishoudens.

6. Validatie. Validatie is kwantitatief en gebeurt door vergelijking met bekende marges. Generalized raking toont betere prestaties dan bestaande methoden qua rekentijd en aansluiting op inputdata. Daarnaast is een empirische toets gedaan met behulp van synthetische populaties voor Zwitserland.

7. Openbaarheid. De gebruikte software is open source en beschikbaar gesteld in R-packages. Ook de datastromen zijn transparant beschreven, met nadruk op reproduceerbaarheid. De volledige toolchain is publiek en herbruikbaar.

8. Toepasbaarheid voor SIVMO. De methode is goed toepasbaar binnen SIVMO-contexten, vooral voor grootschalige toepassingen met veel attributen en hiërarchische relaties. Generalized raking biedt een goede statistische basis en is bruikbaar met reguliere microdata en marges. Kanttekening is dat huishoudsteekproeven met volledige structuur vereist zijn, de beschikbaarheid daarvan kan beperkend zijn.

Müller, K. & Kay W. Axhausen. 2010.
"*Population Synthesis for Microsimulation: State of the Art*." Paper presented at the Swiss Transport Research Conference (STRC), ETH Zürich, August 2010. https://doi.org/10.3929/ethz-a-006127782. eth-1623-01.pdf & Mueller.pdf.

Samenvatting

Het paper bespreekt de stand van zaken in populatiesynthese voor agent-based microsimulatie in transport- en ruimtelijke planning. De auteurs vergelijken zes synthesizers (PopGen, PopSynWin, ILUTE, FSUMTS, ALBATROSS en CEMDAP) die elk een fitting- en een allocatiestap uitvoeren. In de fittingfase worden sampledata aangepast aan marginale distributies met behulp van technieken als Iterative Proportional Fitting (IPF) of Iterative Proportional Updating (IPU). Diverse technische uitdagingen worden besproken, zoals het 'zero-cell'-probleem en geheugenlimieten bij hoge categorisatiedetails. Sommige synthesizers passen automatische categoriereductie of lijstgebaseerde opslag toe om deze problemen te beperken.

In de allocatiefase worden uit de gewogen steekproef huishoudens geselecteerd, vaak via herhaald willekeurig trekken (al dan niet met aanpassing van selectiekansen). Alternatieve methoden, zoals deterministische selectie of conditional Monte Carlo, worden ook besproken. De auteurs signaleren dat deze stap nog weinig theoretisch onderbouwd is en mogelijk bias introduceert.

Tot slot pleiten de auteurs voor de ontwikkeling van een generiek, uitbreidbaar open-source softwareframework voor populatiesynthese. Gezien de diversiteit aan datastructuren en toepassingsdomeinen is een universele synthesizer onwaarschijnlijk, maar herbruikbare modules voor veelvoorkomende taken zouden de praktische toepasbaarheid van deze technieken sterk vergroten.

Analyse

1. Methode. De studie is een literatuuroverzicht waarin zes bestaande populatiesynthesizers voor microsimulatie worden vergeleken. De nadruk ligt op technieken die gebruikmaken van Iterative Proportional Fitting (IPF) of varianten daarvan, zoals Iterative Proportional Updating (IPU). De fitting- en allocatiefasen van elk systeem worden systematisch behandeld. De aanpak is beschrijvend en technisch gedetailleerd, maar bevat geen nieuw algoritme of empirische validatie op eigen data.

2. Inputdata. Er worden geen eigen inputdata gebruikt; het betreft een review. Wel wordt besproken dat populatiesynthese meestal steunt op twee soorten input: een disaggregated sample (zoals een microcensus) en aggregated marginals (zoals bevolkingsstatistieken per zone). Voor sommige tools wordt verwezen naar toepassingen met data uit o.a. Zwitserland, Canada en de VS.

3. Populatiekenmerken. De bespreking richt zich op huishoudens en personen, met controle-attributen zoals leeftijd, geslacht, inkomen en huishoudtype. De synthesizers verschillen in de mate waarin ze personen- en huishoudkenmerken gelijktijdig kunnen verwerken. Bijzondere aandacht gaat uit naar hiërarchische structuren (persoon binnen huishouden) en technieken om deze consistent te synthetiseren.

4. Schaal. Het schaalniveau varieert per synthesizer, van stedelijk (bijv. Toronto, Dallas) tot nationaal (bijv. Zwitserland). De synthese gebeurt vaak op zoneniveau, soms met een regionale hiërarchie. Er is geen expliciete focus op prognoses; de nadruk ligt op basisjaren van simulatiesystemen.

5. Output. De output van de besproken synthesizers is een synthetische populatie van huishoudens en personen, afgestemd op opgelegde marges. Hoewel de nauwkeurigheid van de populaties wordt besproken, worden kwantitatieve prestatiecijfers nauwelijks genoemd. Validatie gebeurt vooral in de originele studies, niet in dit overzicht.

6. Validatie. De auteurs voeren zelf geen validatie uit. Wel wordt gewezen op methoden die door andere auteurs zijn gebruikt (bijv. backcasting bij FSUMTS of goodness-of-fit-maatstaven). Validatie wordt als noodzakelijk beschouwd, vooral voor de allocatiefase waar bias kan ontstaan.

7. Openbaarheid. De besproken synthesizers verschillen in openbaarheid: sommige (zoals PopGen en PopSynWin) zijn als standalone software beschikbaar, anderen zijn ingebed in bredere simulatieplatforms. De auteurs pleiten voor een open, generiek softwareframework.

8. Toepasbaarheid voor SIVMO. De studie is waardevol als overzicht van bestaande methoden en hun technische eigenschappen. Het artikel helpt bij het selecteren of verbeteren van bestaande IPF-gebaseerde routines. De nadruk op flexibiliteit, geheugenverbruik en hiërarchieën sluit aan bij SIVMO-behoeften. ALBATROSS is genoemd als synthesizer.

Rahman, Md. Nobinur & Mahmudur Rahman Fatmi. 2023.
"*Population Synthesis Accommodating Heterogeneity: A Bayesian Network and Generalized Raking Technique*." Transportation Research Record 2677 (6): 41–57. https://doi.org/10.1177/03611981221144289 rahman-fatmi-2023-population-synthesis-accommodating-heterogeneity-a-bayesian-network-and-generalized-raking-technique.pdf

Samenvatting

Deze studie stelt een nieuwe methode voor om synthetische populaties te genereren voor agent-based microsimulaties in stedelijke modellen. De aanpak combineert een Bayesian Network (BN) met een Generalized Raking (GR) techniek. Het doel is om heterogeniteit in huishoudens en individuen beter te modelleren, ontbrekende data in de steekproef te behandelen en een realistische populatie te reconstrueren. De BN wordt gebruikt om een populatiepool te creëren op basis van een microsample, terwijl GR de marginaal bekende totalen op fijnmazige geografische schaal corrigeert. Deze methode wordt toegepast op de regio Okanagan in British Columbia, Canada, met als resultaat een 100% synthetische populatie op disseminatiegebiedniveau.

De auteurs hanteren een gedetailleerde differentiatie tussen huishoudtypes, zoals alleenstaanden, koppels met en zonder kinderen, eenoudergezinnen en overige huishoudens. Binnen elk huishouden wordt een hiërarchische structuur gemodelleerd op basis van leeftijd en inkomen, waarmee de rolverdeling tussen referentiepersoon, partner en kinderen wordt gerepresenteerd. De Bayesian Networks worden opgebouwd via een hybride aanpak van expertkennis en machine learning, waarbij irreële relaties (zoals leeftijd beïnvloedt geslacht) expliciet worden uitgesloten. Vervolgens wordt met forward sampling een grote populatie gegenereerd uit de geleerde netwerken.

De gegenereerde populatie vertoont sterke overeenkomsten met zowel de oorspronkelijke steekproef (PUMF) als met de aggregaatgegevens van de census, op regionaal en DA-niveau. De resultaten tonen aan dat het model nauwkeurige marges en logische variabelenrelaties weet te behouden, hoewel er enige afwijkingen zijn bij attributen met veel categorieën. De combinatie van BN en GR blijkt schaalbaar, accuraat en robuust in het modelleren van sociaal-demografische diversiteit.

Tot slot wordt deze methode geïntegreerd in een breder stedelijk simulatiemodel dat landgebruik, energieverbruik en mobiliteitsgedrag simuleert. De auteurs bevelen aan om verder onderzoek te doen naar interne heterogeniteit binnen huishoudtypes en de toepasbaarheid van deze aanpak op andere datasets. De voorgestelde aanpak biedt daarmee een waardevolle bijdrage aan de populatiesynthesepraktijk binnen ruimtelijke en transportmodellen.

Analyse

1. Methode. De auteurs combineren twee technieken: een *Bayesian Network* (BN) voor het genereren van een synthetische populatiepool, en *Generalized Raking* (GR) voor het afstemmen op controle totalen. De BN wordt data-gedreven geleerd en houdt rekening met hiërarchische huishoudstructuren en missing data via een SEM-

algoritme. GR wordt toegepast als postprocessingstap om de gegenereerde data te laten aansluiten op marginaal bekende verdelingen op gebiedniveau.

2. Inputdata. De methode gebruikt de *Public Use Microdata File* (PUMF) van de Canadese census 2016 (voor BC buiten Vancouver CMA) als microsample. De control totals komen uit dezelfde census en gelden op het niveau van *dissemination areas* (DA's), met gemiddeld 400–700 inwoners. Inputdata omvatten 7 persoonskenmerken en 6 huishoudkenmerken.

3. Kenmerken. De synthese omvat zowel persoons- als huishoudkenmerken, waaronder leeftijd, geslacht, burgerlijke staat, opleiding, inkomen, beroep, huishoudgrootte, woningtype en bouwjaar. De aanpak modelleert afhankelijkheden en heterogeniteit tussen huishoudtypes, bijvoorbeeld het verband tussen leeftijd en gezinssamenstelling.

4. Schaal. Het model werkt op het niveau van *dissemination areas*—de kleinst beschikbare ruimtelijke eenheden in de Canadese census. De case study bestrijkt de regio Okanagan, met 293 DA's, ruim 91.000 huishoudens en 218.000 personen. De schaal is dus zowel micro-geografisch als volledig populatiedekkend.

5. Output. De output is een 100% synthetische populatie van huishoudens en personen op DA-niveau, inclusief toewijzing van sociaal-demografische kenmerken en interne huishoudstructuur. Ook worden onderlinge relaties tussen variabelen behouden. Validatie gebeurt via marginale verdelingen, joint distributions en Cramér's V-waarden.

6. Validatie. De auteurs vergelijken marges van de synthetische populatie met censusdata op regionaal en DA-niveau. De verschillen blijven meestal binnen ±1%, behalve bij attributen met veel categorieën (zoals inkomen). Verder worden de joint distributions beoordeeld en wordt visuele en statistische validatie toegepast (o.a. R^2 en Cramér's V).

7. Openbaarheid. De paper geeft een volledig methodologisch overzicht, inclusief gebruikte algoritmen (BN, SEM, GR, TRS). De gebruikte R-pakketten worden genoemd (zoals bnlearn, mlfit). De code zelf is echter niet expliciet gedeeld of publiek beschikbaar gesteld binnen de publicatie.

8. Toepasbaarheid voor SIVMO. De aanpak is toepasbaar voor SIVMO, zeker vanwege de focus op heterogeniteit, schaalbaarheid en reproduceerbaarheid. Het model is geschikt voor agent-based simulaties en bruikbaar bij beleidsscenario's waar huishoudstructuur, diversiteit en microdata belangrijk zijn. Een kanttekening is dat het model sterk leunt op BN-expertise en relatief intensieve data-voorbereiding.

Rich, Jeppe. 2018.
"*Large-Scale Spatial Population Synthesis for Denmark.*" European Transport Research Review 10 (63). https://doi.org/10.1186/s12544-018-0336-2 . s12544-018-0336-2.pdf.

Samenvatting

In dit artikel beschrijft Jeppe Rich een grootschalige aanpak voor populatie synthese binnen het Deense nationale verkeersmodel (DNTM). De studie richt zich op het modelleren van de gehele Deense bevolking op fijnmazig geografisch en sociaal niveau, als input voor agent-based transportmodellen. De populatiesynthese omvat drie hoofdfasen: harmonisatie van populatiedoelen, matrixfitting met behulp van Iterative Proportional Fitting (IPF), en microsimulatie voor huishoudens.

De harmonisatiefase zorgt ervoor dat alle inputdoelen (zoals leeftijds- en inkomensverdelingen) intern consistent zijn, ook wanneer ze door gebruikers handmatig worden aangepast in scenario's. Dit gebeurt via een hiërarchische rangorde waarbij de meest betrouwbare gegevens als referentie worden gebruikt.

De matrixfitting gebeurt in twee stappen. Eerst worden harmonische doelen gebruikt om een mastertabel van individuen te genereren via IPF (vergelijkbaar met een maximum entropy-aanpak). In een tweede stap kunnen aanvullende doelen op een gedetailleerder geografisch niveau worden toegevoegd. De eindmatrix bevat ruim 17 miljoen cellen en beschrijft sociaaleconomische kenmerken van individuen per zone.

In de simulatiefase worden de prototypische individuen vertaald naar micro-agents en toegewezen aan huishoudens. Deze stap omvat 'spouse matching' (koppelen van partners op basis van sociaaleconomische kenmerken) en het toewijzen van kinderen aan huishoudens, waarbij gebruik wordt gemaakt van marginale verdelingstabellen gebaseerd op registerdata. Omdat deze toewijzingen deels stochastisch zijn, wordt een herweging toegepast om consistentie met de doelen te behouden.

Validatie gebeurt via twee sporen: 1) analyse van de impact van sampling noise op transportresultaten bij herhaalde runs (20 simulaties), en 2) vergelijking van voorspelde bevolkingsaantallen tussen 2010 en 2015 met geobserveerde data. De variatie in modeloutput als gevolg van simulatie-ruis blijkt zeer beperkt (<0,03% op geaggregeerd niveau). De voorspellingen op subzone-niveau laten een gemiddelde afwijking van ca. 3,5% over vijf jaar zien (0,7% per jaar).

De studie concludeert dat een gedetailleerde, meertrapse aanpak van populatiesynthese noodzakelijk is voor betrouwbare modelresultaten, en onderstreept het belang van harmonisatie en huishoud-algoritmen. Voor toekomstig onderzoek wordt gewezen op de behoefte aan betere validatiemethoden, dynamische modellen voor huishoudvorming, robuustere forecast-doelen, en reconstructietechnieken voor de inputdata.

Analyse

1. Methode. De paper beschrijft een hybride benadering voor populatiesynthese, bestaande uit: (i) harmonisatie van doelen, (ii) matrix-fitting met Iterative Proportional Fitting (IPF) en cross-entropieoptimalisatie, en (iii) een

huishoudensimulatie met stochastische toewijzing van individuen tot huishoudens. Daarnaast wordt in een tweede fittingronde zone-specifieke correctie mogelijk gemaakt. De aanpak is deterministisch tot en met de matrix-fitting en introduceert randomness pas bij de huishoudvorming.

2. Inputdata. De input bestaat uit gedetailleerde microdata van de Deense bevolkingsregisters, met variabelen zoals leeftijd, geslacht, inkomen, arbeidsstatus, gezinsstructuur, aantal kinderen en woonzone. De data is zeer volledig en betrouwbaar, mede doordat deze gebruikt wordt voor belastingheffing. Populatiedoelen (constraints) zijn beschikbaar op gemeenteniveau (L0) en subgemeentelijk niveau (L2).

3. Kenmerken. De synthetische populatie is zeer gedetailleerd op individueel niveau (7 dimensies) en is geografisch gespecificeerd tot op L2-zone (907 zones). Naast individuen bevat het model expliciete huishoudens, inclusief partner- en kindtoewijzing, met inachtneming van sociale correlaties (zoals leeftijdsverschillen tussen partners).

4. Schaal. De aanpak dekt de volledige Deense bevolking en bestrijkt alle gemeenten en subgemeentelijke zones. De toepassing richt zich op zowel een basisjaar (2010) als een forecast voor 2015 en een doorkijk naar 2030, al neemt de onzekerheid toe.

5. Output. De output is een gesynthetiseerde microdata-populatie met individuen en huishoudens, die geschikt is als input voor agent-based transportmodellen. Het model produceert ook geaggregeerde tabellen en maakt validatie op zone- en leeftijdsniveau mogelijk. Validatie betreft o.a. voorspellingen van vervoersoutput (trips, mileage) en populatieaantallen op subzone-niveau.

6. Validatie. Er wordt op twee manieren gevalideerd: (i) gevoeligheid voor ‘random seed’ in de huishoudensimulatie (20 runs), en (ii) vergelijking van de voorspelde populatie in 2015 met werkelijke gegevens. De eerste toont minimale ruis (<0.03% op geaggregeerd niveau), de tweede een gemiddelde afwijking van 0.7% per jaar op subzone-niveau.

7. Openbaarheid. De methodologische beschrijving is volledig openbaar. De onderliggende data (Deense registerdata) is deels publiek beschikbaar, maar de gedetailleerde microdata vereist toegang via Statistics Denmark. De code of softwaretool zelf wordt niet gedeeld. De aanpak is reproduceerbaar mits toegang tot soortgelijke data.

8. Toepasbaarheid voor SIVMO. De aanpak is technisch geschikt voor toepassing binnen SIVMO. De uitgebreide aandacht voor harmonisatie, huishoudens en subgemeentelijke zones sluit goed aan bij de Nederlandse situatie. Het model is schaalbaar en geschikt voor beleidsrelevante analyses zoals scenario’s, sociale spreiding en ruimtelijke effecten. De household- en spousematchingmodellen zijn relevant voor AABM-toepassingen.

Rich, Jeppe, Gunnar Flötteröd, Sergio Garrido & Francisco Pereira. 2019.
Review of Population Synthesis Methodologies. Paper presented at the hEART 2019 conference. Department of Management Engineering, Technical University of Denmark. hEART_2019_paper_122.pdf

Samenvatting

Deze paper biedt een systematisch overzicht van bestaande methoden voor populatiesynthese, met bijzondere aandacht voor de toepassing in transport-modellen. De auteurs bepleiten een heldere afbakening van het veld, omdat populatiesynthese zich onafhankelijk heeft ontwikkeld in diverse onderzoeks-domeinen zoals transportvraagmodellering, sociale geografie en ruimtelijke simulaties. Het doel van populatiesynthese is het genereren van een synthetische populatie van individuen (en eventueel huishoudens) die statistisch representatief is voor een doelpopulatie binnen een bepaald geografisch en socio-demografisch kader.

De paper onderscheidt twee hoofdklassen van methoden:
1. Deterministische methoden, zoals matrixfitting en sample expansion, die uitgaan van de volledige betrouwbaarheid van de invoersample;

2. Probabilistische methoden, die uitgaan van de kansverdeling waaruit een sample een realisatie is.

Binnen de probabilistische aanpak onderscheiden de auteurs *data-based resampling* (resampling uit een dataset) en *model-based resampling* (sampling uit een geschat model). De auteurs benadrukken dat model-based resampling nog beperkt wordt toegepast, maar veelbelovend is vanwege de potentie om complexe interacties en correlaties expliciet te modelleren.

De paper biedt daarnaast een overzicht van praktische uitdagingen zoals ‘structural zeros’, schaalbaarheid, beperkingen in constraints, diversiteit, en betrouwbaarheid van voorspellingen. Er wordt gepleit voor meer aandacht voor validatie, integratie van big data en koppeling met downstream transportmodellen. De auteurs voorzien dat het vakgebied snel zal verschuiven richting meer geavanceerde probabilistische technieken, mede onder invloed van ontwikkelingen in machine learning.

Analyse

1. Methode. De paper biedt een conceptuele en systematische classificatie van populatiesynthese methoden. Er wordt onderscheid gemaakt tussen deterministische methoden (zoals matrixfitting en sample expansion) en probabilistische methoden (waaronder data-based en model-based resampling). Het artikel bespreekt de onderliggende aannames, de technische uitvoering en de implicaties van beide benaderingen, zonder zelf een specifieke methode toe te passen.

2. Inputdata. Omdat het een review betreft, worden geen concrete datasets gebruikt. Wel wordt besproken welke typen data doorgaans worden ingezet: microdata (zoals survey- of registerdata), marges (zoals uit censusdata), en eventueel externe modeluitkomsten (bijvoorbeeld uit demografische of landgebruikmodellen). De paper bespreekt ook de rol van big data als opkomende bron.

3. Kenmerken. De paper beschrijft eigenschappen die wenselijk zijn in synthetische populaties, zoals statistische representativiteit, variatie, consistentie met marges, en mogelijkheid tot huishoudstructuurvorming. Er is aandacht voor zowel individuele als huishoudkenmerken, maar geen specifieke uitvoer of voorbeeldpopulatie.

4. Schaal. De methoden zijn breed toepasbaar, van nationale tot wijkniveaus. De auteurs wijzen op het belang van fijnmazige toepassing in stedelijke gebieden, vooral bij beleid over lopen, fietsen of sociale ongelijkheid. Er wordt geen specifieke tijdshorizon gesimuleerd.

5. Output. Er is geen gegenereerde output in deze studie. De besproken methoden leveren doorgaans gesynthetiseerde individuen en huishoudens, soms met extra structuur (zoals sociaal-economische aspecten, leeftijdsverdeling of ruimtelijke spreiding). De auteurs wijzen op het belang van consistentie met inputmarges.

6. Validatie. Validatie wordt besproken als belangrijk, maar lastig onderdeel. De auteurs noemen matching met marges, structurele validatie (bijv. geen causale omkeringen) en tests op diversiteit als mogelijke invalshoeken. Ze wijzen op problemen bij stochasticiteit, structurele nullen en schaalproblemen.

7. Openbaarheid.
De paper bespreekt meerdere publieke en gesloten methoden, maar levert zelf geen code. Wel verwijst men naar bekende technieken (bijv. IPF, Bayesian networks, deep generative models) die deels beschikbaar zijn in R of Python-pakketten. De indeling helpt bij de reproduceerbaarheid van eerdere studies.

8. Toepasbaarheid voor SIVMO. De paper is relevant voor SIVMO: het biedt een breed en systematisch overzicht van beschikbare methoden, met nadruk op hun (on)geschiktheid voor detailniveau, schaalbaarheid en heterogeniteit. Vooral de indeling in deterministisch vs probabilistisch en de onderverdeling binnen de laatste groep is bruikbaar voor een methodologische keuze of afwegingskader.

Saadi, Ismaïl, Hamed Eftekhar, Jacques Teller & Mario Cools. 2016.
"*Investigating the Scalability in Population Synthesis: A Comparative Approach*."
Transportation Planning and Technology 39(6): 569–591.
https://orbi.uliege.be/bitstream/2268/229325/1/_system_appendPDF_proof_hi%20%281%29.pdf. _system_appendPDF_proof_hi (1).pdf

Samenvatting

Deze studie onderzoekt de invloed van schaalbaarheid op de nauwkeurigheid van synthetische populaties. Daarbij worden twee benaderingen vergeleken: de klassieke fitting-based methode (Iterative Proportional Fitting, IPF) en een generation-based methode gebaseerd op een Monte Carlo Markov Chain (MCMC)-algoritme, ook bekend als Gibbs Sampling. Het onderzoek is gericht op het effect van het aantal gesynthetiseerde kenmerken (2–5) en de steekproefgrootte (10%, 25%, 50%).

Het artikel biedt een uitgebreide bespreking van bestaande technieken zoals IPF, IPU, combinatorische optimalisatie, Bayesian Networks en Hidden Markov Models. De auteurs onderbouwen dat simulatie-gebaseerde methoden potentie hebben om accuraat heterogene populaties te genereren, zelfs als bepaalde combinatie-kenmerken ontbreken in de inputdata. De IPF-methode loopt tegen beperkingen aan wanneer het aantal kenmerken toeneemt, vanwege de snelle groei van de k-wegcontingentietabel (curse of dimensionality).

Aan de hand van verschillende validatiemaatstaven (RMSE, MAE en SRMSE) tonen de auteurs dat de keuze van evaluatiemetriek belangrijk is. Terwijl RMSE en MAE een daling van de fout suggereren bij meer kenmerken, laat SRMSE een stijgende trend zien – wat realistischer is bij hogere complexiteit. Hierdoor concluderen de auteurs dat SRMSE beter geschikt is voor het meten van schaalbaarheidseffecten.

De analyse laat zien dat de MCMC-benadering robuuster is voor toenemende aantallen attributen (hogere schaalbaarheid), terwijl IPF minder gevoelig is voor dalende steekproefgroottes. De studie benadrukt dat de keuze van validatiemetriek belangrijk is: RMSE en MAE geven andere trends dan SRMSE. De auteurs concluderen dat bij hogere schaalniveaus de generation-based methode significant beter presteert dan IPF. Het onderzoek vult een leemte in de literatuur door beide filosofieën direct te vergelijken onder variërende schaalcondities.

Analyse

1. Methode. De studie vergelijkt twee populatiesynthese-methoden: Iterative Proportional Fitting (IPF) als fitting-based methode en een simulatie-gebaseerde methode gebaseerd op een Gibbs Sampler (Markov Chain Monte Carlo). Het doel is het analyseren van de effecten van schaalbaarheid (aantal kenmerken) en steekproefgrootte op de nauwkeurigheid van synthetische populaties. De vergelijking is empirisch van aard en richt zich op RMSE, MAE en SRMSE als prestatie-indicatoren.

2. Inputdata. De analyse maakt gebruik van een Belgische workforce survey uit 2013 met 30.700 observaties. Vijf variabelen worden gesynthetiseerd: leeftijd, opleiding, geslacht, beroep en provincie. De steekproefgroottes variëren (10%, 25%, 50%). Voor

IPF worden marges én microdata gebruikt; voor de simulatiebenadering alleen de microdata (PUMS) om conditionele verdelingen te schatten.

3. Kenmerken. De gegenereerde populaties bevatten individuele kenmerken, zonder expliciete huishoudstructuren. De simulatiebenadering blijkt beter in staat om heterogeniteit te behouden en nieuwe combinaties van attributen te genereren die niet aanwezig waren in de inputdata, wat belangrijk is voor realistische agent-based modellen.

4. Schaal. De methoden worden getest voor het synthetiseren van 2 tot 5 kenmerken, wat schaalbaarheid in de zin van dimensionaliteit representeert. De studie laat zien dat de simulatie-gebaseerde methode stabieler presteert bij hogere dimensies, terwijl IPF daar sterk in fout toeneemt. De analyse heeft geen ruimtelijke schaalvariatie of tijdshorizon.

5. Output. De output bestaat uit synthetische populaties per scenario (per combinatie van steekproefgrootte en aantal kenmerken), gevalideerd tegen de volledige originele dataset. Voor beide methoden worden foutmaten per scenario gerapporteerd, met aandacht voor het aantal cellen in de k-wegcontingentietabel.

6. Validatie. Validatie gebeurt kwantitatief via RMSE, MAE en SRMSE. De studie toont dat SRMSE geschikter is voor vergelijking bij toenemende schaalgrootte, terwijl RMSE de fout onderschat bij grote tabellen. De simulatiebenadering blijkt minder gevoelig voor fouten bij meer dimensies of kleinere steekproeven.

7. Openbaarheid. De gebruikte methoden (IPF en Gibbs Sampling) zijn algemeen bekend en publiek beschikbaar. IPF is geïmplementeerd met een R-pakket (‘mipfp’). De simulatie-aanpak is conceptueel beschreven met MNL-modellen en Gibbs Sampler, maar er is geen broncode meegeleverd.

8. Toepasbaarheid voor SIVMO. Deze studie is belangrijk voor SIVMO, vooral voor het selecteren van schaalbare synthese-methoden bij veel kenmerken of beperkte data. De analyse benadrukt de voordelen van probabilistische generatie (zoals MCMC) boven traditionele fitting-methoden bij toenemende complexiteit. De resultaten ondersteunen gefundeerde keuzes bij het ontwerpen van gedetailleerde agent-based modellen.

Significance. 2020.
Toetsingskader Nieuw Nationaal Personenautoparkmodel. Memo aan Konstanze Winter, Remko Smit en Jordy van Meerkerk, 10 november 2020. Den Haag: Significance. M02 - Toetsingskader v04.pdf.

Samenvatting

De notitie van Significance beschrijft het voorgestelde toetsingskader voor een nieuw nationaal personenautoparkmodel. Dit kader moet dienen als leidraad om de werking, kwaliteit en plausibiliteit van het model te beoordelen. De toetsing wordt gefaseerd opgebouwd, beginnend met het vaststellen van streefwaardes per criterium. Die streefwaardes zijn geen harde eisen, maar vormen een signaalfunctie: bij afwijking moet een plausibele verklaring worden gegeven. Het uitgangspunt is vergelijkbaar met de aanpak in het GM4-project, waarbij ook aandacht wordt besteed aan logica, significantie en modeltechnische kwaliteit (zoals log-likelihood).

Het kader omvat zeven beoordelingsaspecten: reproductie van het basisjaar, zeer-kortetermijnprognose (1 jaar), kortetermijnprognose (2–4 jaar), middellangetermijnprognose (10 jaar), langetermijnprognose (20–40 jaar), gevoeligheidsanalyse voor kosten en inkomensveranderingen, en de geschiktheid van het model om beleidsmaatregelen te simuleren. Voor elk aspect zijn indicatoren en bijbehorende streefwaardes uitgewerkt, bijvoorbeeld afwijkingspercentages tussen waargenomen en voorspelde autoaantallen of uitstoot.

Voor de gevoeligheidsanalyse worden elasticiteiten bepaald ten opzichte van vaste, variabele en brandstofkosten, alsook van inkomen. De resultaten worden vergeleken met waarden uit de literatuur. Verwachte bandbreedtes zijn gebaseerd op eerder onderzoek, zoals dat van de Jong & van de Riet (2008) en Significance (2009). Het doel is te toetsen of het model zich ‘realistisch’ gedraagt bij kosten- of inkomenswijzigingen.

Tot slot wordt onderzocht of het model in staat is om waargenomen effecten van beleidsmaatregelen (zoals wijziging van bpm, accijnzen of bijtelling) correct te reproduceren. Als waarnemingen ontbreken, worden vergelijkingen gemaakt met bestaande modellen zoals DYNAMO en CarbonTax. Dit onderdeel vereist afstemming met de opdrachtgever over welke beleidsmaatregelen worden meegenomen.

Analyse

1. Methode. De notitie beschrijft een toetsingskader om de werking en plausibiliteit van een nieuw autoparkmodel (SPARK) systematisch te beoordelen. De aanpak combineert validatie tegen waarnemingen, vergelijking met andere modellen (zoals DYNAMO en CarbonTax), en gevoeligheidsanalyses via elasticiteiten. De methode is niet stochastisch of simulatief op zich, maar vormt een beoordelingsraamwerk waarin meerdere validatieniveaus zijn vastgelegd, oplopend van basisjaarreproductie tot langetermijnprognoses.

2. Inputdata. Er wordt gewerkt met realisatiegegevens over het Nederlandse autopark (zoals autobezit, brandstofsoorten, bouwjaar, emissies en gebruik) voor het basisjaar 2018 en prognosejaren tot 2060. Daarnaast wordt gebruikgemaakt van microdata

(huishoudkenmerken), statistieken uit waarnemingen (zoals CBS, emissies), en resultaten uit andere modellen (DYNAMO, KOTERPA, CarbonTax). Voor gevoeligheids-analyses worden ook literatuurwaarden van elasticiteiten als input gehanteerd.

3. Kenmerken. Het toetsingskader vereist dat het model betrouwbare uitkomsten genereert voor diverse kenmerken: autobezit per huishouden, lease- vs privéauto's, bouwjaarverdeling, gereden kilometers, brandstofgebruik, CO_2- en NOx-uitstoot, fiscale opbrengsten, en gedrag bij beleidswijzigingen. Ook veranderingen per inkomensklasse en leeftijdscategorie worden getoetst, waarmee het model zowel sociaaleconomische als technologische dimensies moet reproduceren.

4. Schaal. Het model en het toetsingskader zijn opgezet voor nationaal gebruik, met nadruk op lange termijn (tot 2060). Tegelijkertijd wordt gedetailleerde informatie op huishoudniveau gebruikt, inclusief segmentaties naar inkomensklassen, leeftijds-categorieën en brandstoftypen. Het model werkt op jaarbasis met focus op structurele trends, korte termijnveranderingen en beleidsgevoeligheid.

5. Output. Het toetsingskader specificeert gedetailleerde kwantitatieve output: onder andere aantallen auto's (lease/privé), verdelingen naar brandstof, emissies, gereden kilometers, en belastingopbrengsten. Voor prognoses worden mutaties ten opzichte van basisjaar vergeleken met waarnemingen en andere modeluitkomsten. Ook worden elasticiteiten berekend als output van gevoeligheidssimulaties.

6. Validatie. Validatie is het centrale doel van het toetsingskader. Er wordt gevalideerd op meerdere niveaus: basisjaarreproductie, zeer-korte termijn (1 jaar), korte, middellange en lange termijn (tot 40 jaar), gevoeligheid voor kosten/inkomen, en gedrag onder beleidswijzigingen. Voor elk aspect zijn streefwaardes of plausibiliteitscriteria gedefinieerd. Er is ook aandacht voor modellogica, significantie en stabiliteit.

7. Openbaarheid. De notitie zelf is intern en niet publiek gedeeld. Er is geen vermelding van open source componenten of vrij beschikbare software. Wel wordt verwezen naar bestaande modellen en gepubliceerde literatuur. De reproduceer-baarheid van het kader hangt af van toegang tot model en data, die hier niet zijn meegeleverd.

8. Toepasbaarheid voor SIVMO. De systematiek van dit toetsingskader is relevant voor SIVMO. Het kader biedt een concreet beoordelingsraamwerk om de prestaties van modellen voor voertuigen en gedrag over tijd te evalueren, inclusief gevoeligheden en beleidsimpact. De nadruk op validatie, transparantie van aannames, en benchmarking maakt het toepasbaar voor toetsing van populatiesynthese en gedragsmodellen binnen mobiliteitsonderzoek.

Significance. 2021.

Backcast LMS: Vergelijking prognose en waargenomen ontwikkeling. Rapportnummer 21027. Den Haag: Significance. 21027-R01 LMS backcast eindrapport.pdf

Samenvatting
Dit rapport beschrijft een zogeheten "backcast"-analyse met het Landelijk Model Systeem (LMS) voor de periode 1995–2014. In plaats van vooruit te modelleren naar de toekomst, wordt hier met behulp van historische inputdata (zoals demografie, economie, infrastructuur en vervoersprijzen) het verleden gereconstrueerd, zodat gemodelleerde verkeersvolumes en herkomsten-bestemmingen vergeleken kunnen worden met geobserveerde uitkomsten.

Doel van de analyse is om te onderzoeken of het LMS (versie 2020) de historische ontwikkeling van vervoer in Nederland adequaat kan reproduceren. Dit geeft inzicht in de structurele prestaties van het model, los van toekomstige onzekerheden. De reconstructie richt zich op verplaatsingsgedrag van personen, gesegmenteerd naar vervoerwijze, motief, tijdstip, regio en sociaaleconomische groep.

Het rapport maakt duidelijk dat er verschillen bestaan tussen gemodelleerde en gerealiseerde verplaatsingen, vooral in stedelijke gebieden en voor fiets en openbaar vervoer. Oorzaken worden gezocht in het model zelf (keuzemodellen, aanbod-representatie) én in de gebruikte invoer (zoals het huishoudensbestand). Hoewel het rapport geen populatiesynthese uitvoert, wordt wel aandacht besteed aan de kwaliteit en samenstelling van de huishouddata als fundamenteel onderdeel van de modelinvoer.

De backcast-analyse draagt bij aan modelvalidatie en mogelijke toekomstige aanpassingen in het LMS. De nadruk ligt op methodologische transparantie en het interpreteren van afwijkingen.

Analyse

1. Methode. Geen populatiesynthese uitgevoerd. Het rapport beschrijft een toepassing van het LMS in backcast-modus: het met historische invoerdata nabouwen van verplaatsingsgedrag in het verleden, ter validatie van het systeem. Er is geen specifieke synthese-algoritme toegepast.

2. Inputdata. Historische input voor de jaren 1995 en 2014, waaronder bevolkingsaantallen, huishoudsamenstelling, economische variabelen (inkomen, werkgelegenheid), netwerken, modal split, en tarieven. Huishouddata komt deels uit CBS en wordt geconverteerd naar LMS-indelingen.

3. Kenmerken. De huishoudensdata bevat leeftijd, huishoudtype, autobezit, inkomen, werk- en schoolstatus. Deze worden door het LMS gebruikt om verplaatsingskansen en modaliteit te bepalen. De kenmerken zijn niet synthetisch gegenereerd, maar overgenomen uit CBS-tellingen en enquêtes.

4. Schaal. Nationaal niveau (heel Nederland), met segmentatie naar zones, landsdelen en stedelijke regio's. Resolutie afgestemd op het LMS-netwerk en -zonering (COROP-niveau of fijnmaziger).

5. Output. Het model levert verplaatsingsstromen per motief, tijdstip, modaliteit en regio. De output wordt vergeleken met waarneemdata (zoals OV-chipkaartdata, verkeersintensiteiten) over dezelfde periode. Er worden verschillen gerapporteerd tussen gemodelleerde en feitelijke volumes.

6. Validatie. Vergelijking van modeloutput met realisaties (observaties) op netwerk-, regionaal en nationaal niveau. Er is geen formele validatie van een synthetische populatie. Validatie betreft het hele LMS-systeem en zijn gedrag over tijd.

7. Openbaarheid.
Het gebruikte LMS is niet open source, maar de rapportage is transparant over veronderstellingen en data-invoer. De analyse is reproduceerbaar voor partijen met toegang tot het LMS.

8. Toepasbaarheid voor SIVMO. Dit rapport bevat geen populatiesynthese en is dus niet direct relevant voor methodevergelijking. Wel illustreert het de rol die huishoudensdata en modelinvoer spelen in transportmodellen. Indien SIVMO het LMS (of vergelijkbare modellen) wil voeden met synthetische populaties, dan is dit rapport relevant als context voor invoereisen en validatieproblemen.

Significance. 2022a.
QUAD en GWI: Resultaten Fase I. Rapportnummer 22050. Den Haag: Significance. 22050 R01 QUAD en GWI - Fase I versie 4.pdf.

Samenvatting

In dit rapport onderzoekt Significance het functioneren van de bestaande QUAD-module voor populatiesynthese binnen het landelijke verkeersmodel LMS, in samenhang met het General Welfare Index (GWI). De aanleiding is dat gebruikers regelmatig signaleren dat bepaalde bevolkingsgroepen of kenmerken in de synthetische populatie van het LMS onder- of oververtegenwoordigd zijn, met als gevolg onnauwkeurigheden in verkeersprognoses.

QUAD is een systeemeigen component van LMS die populatiesynthese uitvoert via een ophogingsmechanisme. Daarbij worden huishoudens gesampeld uit CBS-micro-data, zodanig dat gewichten per populatie-eenheid (zone × groep) overeenkomen met de gewenste totalen in het GWI. Die gewichten zijn echter verouderd, en sluiten niet altijd aan bij de meest actuele gegevens of modellen.

De studie analyseert op systematische wijze hoe het ophogen in QUAD plaatsvindt, welke onderdelen worden beïnvloed door keuzes in het GWI, en waar de belangrijkste gevoeligheden zitten. In een deel van de zones wordt het populatievolume niet gehaald of ontstaan huishoudtypes die in werkelijkheid zeldzaam of onmogelijk zijn. Daarnaast is de afhankelijkheid van de huishoudverdeling in GWI problematisch, vooral voor toekomstscenario's.

Het rapport doet aanbevelingen voor verbetering van de methodiek, waaronder het herzien van het GWI, het integreren van nieuwe CBS-data, en het overwegen van alternatieve synthesemethoden (bijv. IPF of combinatoriële optimalisatie). Ook wordt gesuggereerd om QUAD modulair te maken en te voorzien van transparante documentatie en tests.

Analyse

1. Methode. QUAD gebruikt een ophogingsmethode waarbij microdata (personen en huishoudens) gesampled worden met gewichten per zone–groepcombinatie. Er is geen IPF of optimalisatie toegepast; het algoritme controleert of het gewenste volume wordt gehaald via selectie en ophoging, niet of de interne samenstelling overeenkomt met waargenomen joint distributions.

2. Inputdata. CBS-gegevens vormen de microdata (basispopulatie). Het GWI bevat control totals per zone en type huishouden/persoon, maar is deels gebaseerd op aannames of historische data. GWI bepaalt in sterke mate de uitkomst van de populatiesynthese.

3. Kenmerken. De synthetische populatie waarop het model is toegepast bevat differentiatie naar huishoudsamenstelling, inkomen en demografische kenmerken. De gevoeligheid van het model wordt vooral getoetst aan variatie in de GWI-componenten. In hoeverre individuele kenmerken zoals leeftijd of opleiding invloed hebben op mobiliteit via inkomensontwikkeling wordt expliciet geanalyseerd.

4. Schaal. Nationaal model met zonale resolutie (modelzones LMS). Synthese gebeurt zone voor zone, zonder expliciete koppeling of consistentiecontrole tussen zones. Onbalans of onmogelijke combinaties komen vooral in kleine zones voor.

5. Output. De output is een populatiebestand op huishoud- en persoonsniveau, bedoeld voor gebruik in LMS. Structuur is technisch bruikbaar maar bevat soms contra-intuïtieve combinaties of ontbrekende subgroepen. Kwaliteit hangt sterk af van GWI en samplinglogica.

6. Validatie. Beperkte validatie aanwezig. Er is geen systematische vergelijking tussen QUAD-output en CBS-tellingen of onafhankelijke bronnen. Veel problemen worden pas zichtbaar in latere LMS-output (bijv. onnatuurlijke verdeling van verplaatsingen of autobezit).

7. Openbaarheid. QUAD is geen open tool; het is niet los bruikbaar buiten het LMS. De gebruikte CBS-data zijn niet openbaar.

8. Toepasbaarheid voor SIVMO. QUAD is beperkt toepasbaar als standalone PS voor SIVMO. De methode is onvoldoende robuust voor scenarioanalyse. Voor SIVMO-doeleinden zijn alternatieven als IPF, combinatoriële optimalisatie of deep generative modellen waarschijnlijk beter geschikt. De aanbevelingen in dit rapport kunnen echter wel richting geven aan de eisen voor toekomstige tools.

Significance. 2022b.
QUAD en GWI: Fase Resultaten Fase II. Rapportnummer 22050. Den Haag: Significance. 22050 R02 QUAD en GWI - Fase II.pdf

Samenvatting

Het document beschrijft de tweede fase van het onderzoeksproject naar verbetering van de methoden QUAD en GWI binnen het Landelijk Model Systeem (LMS). In deze fase is het doel om voorgestelde aanpassingen uit fase I door te voeren, te testen en te evalueren. De focus ligt op het exogeen maken van de General Welfare Index (GWI) en het verbeteren van de stabiliteit en consistentie van het QUAD-model bij het genereren van populaties voor verkeersprognoses.

Een belangrijk onderdeel van het onderzoek is de toetsing van het effect van de exogene GWI. In het scenario voor 2040L leidde de oorspronkelijke GWI tot een te lage mobiliteit, ondanks lichte inkomensgroei. Door een aangepaste GWI van 0,9920 toe te passen, wordt een realistischere groei in autokilometers bereikt. Hoewel dit de aansluiting met waarnemingen niet significant verbetert, biedt het wel een beter uitlegbaar verhaal. De nieuwe GWI maakt het mogelijk om scenario's met meer controle te formuleren.

Daarnaast zijn verschillende technische aanpassingen aan het QUAD-model getest, zoals het variëren van het gewicht van onderwijstargets en het gebruik van zonale microdata. Dit leidde tot verbeteringen in de aansluiting van het model op gewenste kenmerken, zoals huishoudtype en opleidingsniveau. De robuustheid van het model nam toe bij bepaalde instellingen, al blijft calibratie noodzakelijk.

Tot slot concludeert het rapport dat de exogene GWI bijdraagt aan betere uitlegbaarheid en bruikbaarheid van het model. De aanpassingen aan QUAD worden als waardevol beschouwd voor het maken van toekomstbestendige prognoses, mits goed onderbouwd en afgestemd op empirische data. Verdere evaluatie en validatie blijven nodig.

Analyse

1. Methode. Het rapport beschrijft een QUAD in combinatie met een exogeen gemaakte index voor algemene welvaart (GWI). Fase II richt zich op het verbeteren van modelgedrag en aansluiting met waargenomen data via technische aanpassingen (zoals het uitschakelen van specifieke targets) en op het herijken van GWI als externe input voor scenariomodellen. De validatie richt zich vooral op de aansluiting met eerdere modelrondes en plausibiliteit van de prognose-uitkomsten.

2. Inputdata. Het model maakt gebruik van microdata op huishoudenniveau (zoals huishoudtype, inkomen, opleidingsniveau, autobezit), aangevuld met zonale gegevens en projecties voor 2040. Verder zijn modeluitkomsten uit het LMS en eerdere QUAD-runs input voor validatie en vergelijking. De exogene GWI is afgeleid uit een kalibratie op gerealiseerde autokilometers en inkomensgroei.

3. Kenmerken. De synthetische populatie omvat huishoudens met kenmerken zoals huishoudtype, inkomen, opleidingsniveau, autobezit en levensfase. De nadruk ligt op

consistentie met marges en plausibiliteit van uitkomsten, maar er is geen expliciete structurering van gezinsverbanden of gedrag. De toevoeging van een externe GWI maakt de populatie-ontwikkeling gevoeliger voor beleidskeuzes.

4. Schaal. Het model is toegepast op landelijk niveau met een uitsplitsing naar zones. Het richt zich op lange termijnprognoses (2040), maar bevat ook tests met kortere horizon (bijv. voor trendherkenning en validatie). De schaal is consistent met strategische beleidsmodellen zoals LMS.

5. Output. Het resultaat is een synthetische populatie voor 2040 met geprojecteerde kenmerken die invoer is voor verkeers- en mobiliteitsmodellen. Belangrijke uitkomsten zijn het aantal voertuigen, gereden kilometers, autobezit en verdelingen naar sociaal-demografische klassen.
6. Validatie. Validatie gebeurt via vergelijking met eerdere prognoses en via aansluiting op bekende marges (zoals de groei in autokilometers in scenario 2040L). Ook worden gevoeligheidsanalyses gedaan door variatie van modelinstellingen en de GWI-waarde. Absolute validatie op waarnemingen ontbreekt, omdat het om toekomstscenario's gaat.

7. Openbaarheid. De gebruikte modellen QUAD en GWI zijn niet publiek beschikbaar, maar worden in opdracht van Rijkswaterstaat en het Kennisinstituut voor Mobiliteitsbeleid onderhouden. De documentatie in dit rapport is beknopt maar informatief; code of datasets worden niet gedeeld.

8. Toepasbaarheid voor SIVMO. De aanpak sluit aan bij SIVMO-doelstellingen vanwege de expliciete aandacht voor scenario's, heterogeniteit en modelcontrole. De exogene GWI maakt het mogelijk om beleidsmatige sturing op welzijn en gedrag systematisch te verwerken. Wel vereist toepassing toegang tot de modellen en kennis van kalibratie.

Significance. 2024a.
Vergelijking QUAD – SigPopu. Intern rapport, versie februari 2024. Den Haag: Significance. Vergelijking QUAD - SigPopu - v5.pdf

Samenvatting

Dit document is een intern rapport van Significance waarin twee population synthesizers worden vergeleken: QUAD, het bestaande instrument in het LMS-systeem, en SigPopu, een alternatieve implementatie gebaseerd op minimalisatie van de Kullback-Leibler-divergentie. De aanleiding is dat de uitkomsten van QUAD instabiel blijken tussen verschillende zichtjaren. Dat kan gevolgen hebben voor doorrekeningen met het LMS-model, omdat de ophoogfactoren die QUAD genereert per jaar sterk verschillen. De centrale vraag is of SigPopu deze instabiliteit kan verminderen en betere modelresultaten oplevert.

Het rapport vergelijkt de twee synthesizers aan de hand van vijf vragen: (1) hoe goed ze voldoen aan de opgelegde marges (targets), (2) hoe dicht ze bij de a-priori populatie blijven, (3) hoe sterk de ophoogfactoren overeenkomen, (4) hoe stabiel die ophoogfactoren zijn over jaren heen, en (5) wat het effect is op een LMS-run.
Er zijn drie runs uitgevoerd: QUAD, SigPopu-A (met dezelfde vrijheidsgraden als QUAD) en SigPopu-B (met individuele ophoogfactoren per huishouden). SigPopu-B behaalt alle targets exact, SigPopu-A vrijwel allemaal (typische afwijking 0,1%) en QUAD mist meerdere targets, vooral bij rijbewijsbezit, inkomen en opleidingsniveau. SigPopu komt ook dichter bij de oorspronkelijke (a-priori) huishoudverdeling dan QUAD, zelfs al optimaliseert SigPopu daar niet op. De standaarddeviatie van de afwijking is bij QUAD ruim vier keer zo groot als bij SigPopu-B. Dit effect wordt waarschijnlijk veroorzaakt door het iteratief op nul zetten van negatieve ophoogfactoren in QUAD.

SigPopu toont bovendien meer stabiliteit over tijd: waar QUAD categorieën in het ene jaar groot en in het andere jaar nul maakt (“mikado-patroon”), blijven de verschillen bij SigPopu-B veel kleiner. Bij toepassing in het LMS-model zijn de nationale uitkomsten in termen van totaal aantal reizen vrij vergelijkbaar (verschillen <1%), maar de verdeling naar motieven en modaliteiten verschilt per cel tot 5%, en de groei tot 2040 wijkt af, vooral bij passagiersautogebruik, fietsen, lopen en e-bike.

De conclusie is dat SigPopu, en vooral de B-variant, betere, stabielere en realistischere populaties oplevert. Aanbevolen wordt om QUAD te vervangen door SigPopu, inclusief benodigde aanpassingen in SES en CARMOD.

Analyse

1. Methode. De studie vergelijkt twee methoden voor populatiesynthese: QUAD (gebaseerd op kwadratische optimalisatie) en SigPopu (gebaseerd op minimalisatie van Kullback-Leibler-divergentie). SigPopu kent twee varianten: A (ophogen per categorie) en B (ophogen per huishouden). De methoden worden getest op hun vermogen om populaties te genereren die zowel voldoen aan targets als vergelijkbaar blijven met de a-prioripopulatie. SigPopu gebruikt een probabilistische benadering zonder negatieve gewichten, terwijl QUAD iteratief werkt en gewichten op nul zet als ze negatief dreigen te worden.

2. Inputdata. Beide methoden gebruiken dezelfde prototype steekproef en dezelfde set van 20 zone-targets en 8 landelijke targets. De vergelijking is uitgevoerd voor twee jaren (2018 en 2040), waarbij de targets betrekking hebben op demografie, opleiding, inkomen en rijbewijsbezit. Voor LMS-doorrekeningen wordt ook gebruikgemaakt van reizen per motief en vervoerwijze.

3. Kenmerken. De uitkomsten zijn populaties met huishoudcategorieën, ophoogfactoren en resulterende sociaaldemografische verdelingen. SigPopu garandeert positieve gewichten en levert populaties die beter aansluiten bij de inputverdeling. Kenmerken als consistentie over tijd, het niet volledig uitsluiten van huishoudcategorieën en betere aansluiting op targets worden als voordelen genoemd.

4. Schaal. De analyse is uitgevoerd op nationaal niveau en voor 1406 zones. Dit toont aan dat beide methoden geschikt zijn voor toepassing op hoge geografische resolutie. SigPopu-B presteert hierbij duidelijk beter in termen van stabiliteit en nauwkeurigheid.

5. Output. Synthetische populatie, ophoogfactoren per huishouden(categorie), afwijkingen t.o.v. targets, afwijkingen t.o.v. a-priori populatie, effecten op LMS-reisuitkomsten (totaal, per motief en modaliteit), en groeicijfers 2018–2040.

6. Validatie. Interne vergelijking van afwijkingen en prestaties op targets. Geen externe validatie, maar wel vergelijking met bekende marges en controle op consistentie in tijd.

7. Openbaarheid. Het document beschrijft SigPopu als gebaseerd op bestaande technieken en als herbruikbare optimalisatieroutine, maar biedt geen publieke code of documentatie. QUAD is een bestaande interne tool. Beide methoden zijn deels ingebouwd in Groeimodel of kunnen dat worden, maar zijn niet openbaar beschikbaar in standaardsoftwarevorm.

8. Toepasbaarheid voor SIVMO. De studie is direct toepasbaar voor SIVMO. Ze laat zien hoe verschillen in optimalisatietechniek kunnen leiden tot afwijkende beleidsindicaties. SigPopu blijkt consistenter en beter in staat om heterogene huishoudens representatief te modelleren. De analyse biedt waardevolle inzichten voor methodologische keuzes bij populatiesynthese in regionale en nationale modellen.

Significance. 2024b.
Actualisatie invoer huishoudsimulator. Memo 24010-M02 v8, 22 mei 2024. Den Haag: Significance. 24010-M02 - Actualisatie invoer huishoudsimulator v08.pdf

Samenvatting

Deze memo beschrijft de actualisatie van de invoerbestanden van de huishoudsimulator SPARK. De invoer betreft zowel de doelvariabelen (targets) als de overgangskansen (probabilities), die zijn geactualiseerd voor de nieuwe WLO III-scenario's: Hoog, Laag en KEV. Hiervoor zijn nieuwe DEMOGRAF- en SEGS-bestanden gebruikt, naast CBS StatLine-data voor de periode 2019–2023.

Voor de targets zijn onder meer sterfte, geboortes, huishoudgroottes, verhuisstromen, sociaaleconomische status, sector, en inkomen per regio en stedelijkheidsklasse geactualiseerd. Daarbij is per variabele een aanpak gekozen die zoveel mogelijk aansluit bij bestaande datastructuren en publicaties, aangevuld met eigen analyses waar nodig. CBS-prognoses en PBL-doorrekeningen zijn gecombineerd met lineaire schaling en correcties om consistentie met nationale totalen te waarborgen.

Voor de overgangskansen is een nieuwe set DEMOGRAF-bestanden als basis genomen. Overlijdens- en geboorteprognoses zijn direct uit deze bestanden gehaald. Transities in huishoudsamenstelling zijn herleid uit samengestelde transitiebestanden, waarbij kansstructuren zijn opgebouwd op basis van individuele gebeurtenissen en gewogen over bevolkingsgroepen. Kansen op verhuizing, sectorwisseling en verandering van sociaaleconomische status zijn gebaseerd op CBS-microdata. De inschatting van pensioenuitstroom is expliciet meegenomen met stapsgewijze overgangsregels gebaseerd op toekomstige AOW-leeftijden.

De memo bevat appendices waarin methoden voor extrapolatie, schaling en clustering zijn toegelicht. Er zijn rekenregels geformuleerd voor de conversie tussen verschillende huishoudposities en scenario's. De simulator blijft gebaseerd op dezelfde structuur, maar is door deze update inhoudelijk op peil gebracht voor toepassing in recente WLO-verkenningen. De bijgevoegde CBS-bronnen en formules geven transparantie over herkomst en bewerking van de data.

Analyse

1. Methode. De methode betreft een deterministische actualisatie van invoerbestanden voor een bestaande huishoudsimulator. Er worden geen stochastische technieken toegepast: de targets en overgangskansen worden vastgesteld op basis van CBS-data en scenario-extrapolaties. Voor ontbrekende waarden zijn lineaire extrapolaties of verdelingen toegepast. De aanpak is vooral input-gestuurd en sluit aan op het SPARK-model.

2. Inputdata. DEMOGRAF-bestanden, SEGS, CBS StatLine, CBS-microdata (2018–2023), PBL-scenario's voor KEV en WLO III. Diverse datasets per huishoudtype, regio, sector en leeftijd.

3. Kenmerken. Disaggregate huishoudsimulatie per individu/huishouden; jaarlijkse actualisatie; scenario-afhankelijk. Expliete overgangskansen per type gebeurtenis en doelgroep.

4. Schaal. Nationaal niveau met uitsplitsing naar LMS-zones, stedelijkheidsklasse en sector. De tijdshorizon loopt van 2019 tot 2060.

5. Output. De output bestaat uit geactualiseerde targetbestanden en overgangskansen per jaar en scenario. Deze worden gebruikt in de SPARK-simulator om gesynthetiseerde huishoudpopulaties te genereren, inclusief attributen als samenstelling, inkomen en werkstatus.

6. Validatie. Targets voor 2019–2023 gevalideerd aan de hand van CBS-realisatie. Waar nodig lineaire benaderingen toegepast om ontbrekende cijfers te reconstrueren.

7. Openbaarheid. Gebruikte microdata van het CBS is alleen intern beschikbaar onder specifieke gebruiksvoorwaarden.

8. Toepasbaarheid voor SIVMO. Relevant als benchmark voor huishoudsimulatie in Nederland. Sluit aan bij SPARK en het bredere LMS-systeem. Methode is reproduceerbaar en transparant.

Wu, Hao & Cheng Lyu. 2024.
Simulation-Based Comparative Analysis of Synthetic Population Generation Methods: A Framework for Travel Diary Validation in Transport Simulation. Chair of Transportation Systems Engineering, TUM School of Engineering and Design, Technische Universität München.
Expose_Simulation_Based_Comparative_Analysis_of_Synthetic_Population_Generation_Methods.pdf

Samenvatting
In dit projectvoorstel presenteren Wu en Lyu een methodologisch kader voor het valideren van synthetische populaties in agent-based verkeerssimulaties. Het centrale doel is het kwantificeren van de overeenkomst tussen synthetische populaties en empirische verplaatsingsdata (reisdagdboeken), met toepassing binnen het MATSim-platform. De auteurs constateren dat bestaande populatiesynthesemethoden zich zelden laten toetsen op het niveau van activiteitspatronen, ondanks de cruciale rol van gedrag in vervoersmodellen.

Het voorstel omvat de ontwikkeling van een wiskundig vergelijkingsframework waarin zowel structurele als verdelingsverschillen tussen echte en gesynthetiseerde verplaatsingsreeksen worden gemeten. Hiervoor introduceren zij twee kerncomponenten:

1. Structurele gelijkenis tussen reisdagdboeken, gebaseerd op meerdere metrics (bijv. duur, volgorde, modaliteit), gecombineerd in een gewogen som.

2. Distributievergelijking, via gecombineerde toepassing van Kullback-Leibler divergence (DKL) en Maximum Mean Discrepancy (MMD).

De implementatie gebeurt in Java en sluit aan op het MATSim-systeem. Reisdatabase en synthetische populaties worden gekoppeld aan dit validatiekader, waarmee foutmarges in syntheseprocessen geobjectiveerd kunnen worden. De methode is bedoeld als aanvulling op bestaande validatieroutines die zich vooral richten op marges of socio-demografische verdelingen.

Het einddoel is om een breed toepasbaar, reproduceerbaar framework te ontwikkelen dat onderzoekers en beleidsmakers ondersteunt bij het beoordelen van de kwaliteit van gesynthetiseerde populaties in transportmodellen. De planning voorziet in een looptijd van zes maanden, met deelstappen voor literatuur, theoretische ontwikkeling, implementatie en validatie.

Analyse
1. Methode. De voorgestelde aanpak is een mathematisch-statistisch validatiekader dat synthetische populaties vergelijkt met reisdagboekdata. Het combineert structurele gelijkenismaten tussen activiteitspatronen met verdelingsvergelijking op populatieniveau. Er wordt gebruikgemaakt van bestaande populatiesynthese methoden en simulatie in MATSim. De aanpak is niet gericht op synthese, maar op toetsing van bestaande synthetische populaties.

2. Inputdata. De methode vereist twee typen data: (1) echte reisdagboekdata van individuen (DR), en (2) synthetisch gegenereerde activiteitspatronen (DS), bijvoorbeeld uit bestaande synthesizers gekoppeld aan MATSim. Deze datasets worden vergeleken via gedefinieerde metrics.

3. Kenmerken. Populaties worden vergeleken op het niveau van individuele activiteitenketens, inclusief activiteitentype, volgorde, locatie, duur en modaliteit. Verdeling over de populatie wordt statistisch geëvalueerd.

4. Schaal. Niet gespecificeerd, maar bedoeld voor gebruik in realistische MATSim-toepassingen. De schaal is afhankelijk van de beschikbaarheid van travel diaries en synthetische output per casus.

5. Output. De output bestaat uit numerieke maatstaven voor de gelijkenis tussen synthetische en echte populaties. Dit omvat metrics zoals een gewogen som van structuurvergelijking en verdelingsvergelijking (o.a. Kullback-Leibler Divergence en Maximum Mean Discrepancy). De resultaten kunnen worden gebruikt om synthesizers te beoordelen.

6. Validatie. Validatie *van* populaties is het centrale doel. Er worden formele vergelijkingsfuncties ontwikkeld en toegepast, inclusief sensitiviteitsanalyse. Validatie gebeurt statistisch en gedragsmatig.

7. Openbaarheid. De methodologische specificatie is publiek. Implementatie gebeurt in Java. MATSim is open source; de integratie met dit platform maakt hergebruik en uitbreiding goed mogelijk.

8. Toepasbaarheid voor SIVMO. Bruikbaar voor SIVMO als validatiekader om gegenereerde populaties objectief te beoordelen op gedragsrealiteit, vooral bij toepassing van agent-based modellen. Hoewel geen synthese-algoritme wordt ontwikkeld, sluit de aanpak goed aan bij kwaliteitsborging en modelvergelijking.

Wu, Hao & Cheng Lyu. 2024.
Tabular Data Imputation for Synthetic Population with Diffusion Models. Chair of Transportation Systems Engineering, TUM School of Engineering and Design, Technische Universität München.
Expose_Tabular_Data_Imputation_for_Synthetic_Population_with_Diffusion_Models.pdf.

Samenvatting
Dit projectvoorstel richt zich op de ontwikkeling van een nieuwe methode voor synthetische populatiesynthese die expliciet is ontworpen om te gaan met ontbrekende waarden in tabulaire datasets. Traditionele technieken zoals IPF of Bayesian Networks gaan meestal uit van volledige datasets of gebruiken simpele imputatietechnieken die bias kunnen introduceren. Het doel van dit onderzoek is om een diffusion model (meer specifiek: TabCSDI – Tabular Conditional Score-based Diffusion Imputation) toe te passen en aan te passen voor populatiesynthese.

Het theoretisch kader combineert twee stromingen in machine learning: (1) deep generatieve modellen zoals VAE voor populatiesynthese, en (2) recente diffusion models voor tabulaire data-imputatie. De innovatieve stap is deze technieken te integreren in één coherent model dat gelijktijdig plausibele populaties kan genereren én ontbrekende data kan imputereren.

De data bestaat uit nationale travel survey-data (inclusief socio-demografische en mobiliteitskenmerken), aangevuld met benchmarkdatasets. Traditionele methoden (IPF, Gibbs sampling, Bayesian Networks) worden als baseline geïmplementeerd. Evaluatie vindt plaats op basis van SRMSE, Pearson correlatie en Cramér's V, met variërende vormen van missingness (MCAR, MAR, MNAR). Er is ook aandacht voor externe validatie via agent-based simulaties.

De verwachte meerwaarden zijn: (1) een robuust raamwerk voor synthese met missing data, (2) empirische vergelijking met klassieke methoden, (3) praktische richtlijnen voor encoding van gemengde variabelen, en (4) inzicht in de impact van imputatiekwaliteit op downstream verkeersmodellen.

Analyse
1. Methode. Het voorstel introduceert een nieuwe aanpak voor populatiesynthese waarbij diffusion models worden ingezet voor het imputeren van ontbrekende waarden in tabulaire data. De methode combineert bestaande technieken uit de machine learning: traditionele synthesemethoden (IPF, Gibbs sampling, Bayesian Networks), deep generative modellen (VAE) en recent ontwikkelde diffusion models (zoals TabCSDI). De opzet is conceptueel innovatief: imputation en synthese worden niet apart maar simultaan behandeld binnen één generatief model.

2. Inputdata. De belangrijkste databron is een nationale reisdagboekdataset, waarin socio-demografische en mobiliteitskenmerken van individuen zijn opgenomen. Daarnaast worden andere benchmarkdatasets gebruikt om de generaliseerbaarheid te testen. De data bevat zowel numerieke als categorische variabelen en bevat doelbewust ontbrekende waarden, om de robuustheid van de methode te kunnen evalueren.

3. Kenmerken. Focust op gemengde attributen en ontbrekende waarden. Encoding technieken worden onderzocht als integraal onderdeel van het modelontwerp.

4. Schaal. Geen expliciete geografische schaal; toepasbaar op nationale datasets. Model moet generaliseerbaar zijn naar andere contexten.

5. Output. De output is een volledig gesynthetiseerde populatie met geïmputeerde gegevens die geschikt is voor gebruik in agent-based simulatiemodellen. Ook worden concrete prestatie-indicatoren gegenereerd die de kwaliteit van de synthetische populatie beschrijven.

6. Validatie. Er is sprake van zowel interne als externe validatie. Intern worden statistische vergelijkingen gemaakt tussen echte en synthetische populaties. Extern wordt de gegenereerde populatie ingezet in transportmodellen om het effect van datakwaliteit op modeluitkomsten te onderzoeken. Daarnaast wordt een gevoeligheidsanalyse uitgevoerd op de gebruikte parameters.

7. Openbaarheid. Onderzoeksmateriaal is deels gebaseerd op publieke modellen en datasets. TabCSDI is beschikbaar als arXiv-preprint; MATSim is open source. Implementatie in Python.

8. Toepasbaarheid voor SIVMO. De voorgestelde methode is potentieel waardevol voor SIVMO, vooral vanwege de aandacht voor datakwaliteit en ontbrekende waarden. De aanpak biedt een technisch robuust alternatief voor IPF en vergelijkbare technieken, met meer flexibiliteit en betere integratie met moderne simulatie-modellen. De reproduceerbaarheid en inzetbaarheid hangen af van beschikbaarheid van code en schaalbaarheid.

Ye, Xin, Karthik C. Konduri, Ram M. Pendyala, Bhargava Sana & Paul Waddell. 2009. *"A Methodology to Match Distributions of Both Household and Person Attributes in the Generation of Synthetic Populations."* Paper gepresenteerd op de 88e Annual Meeting of the Transportation Research Board, Washington, D.C., januari 2009.

Samenvatting

Deze studie introduceert de Iterative Proportional Updating (IPU)-algoritme als een nieuwe methode voor het genereren van synthetische populaties, waarbij zowel huishoud- als persoonskenmerken gelijktijdig worden gematcht aan bekende verdelingen. Traditionele methoden zoals Iterative Proportional Fitting (IPF) houden alleen rekening met huishoudkenmerken, wat leidt tot slechte overeenkomsten op persoonsniveau. Het IPU-algoritme past iteratief de gewichten van huishoudens aan, zodat beide niveaus van kenmerken zo goed mogelijk worden benaderd.

Het algoritme is praktisch toepasbaar, ook voor kleine geografische eenheden zoals blokgroepen, en vereist beperkte rekenkracht. De auteurs testen het IPU-algoritme in Maricopa County (Arizona) met behulp van Amerikaanse Census-data. Ze tonen aan dat het algoritme effectief is in het verminderen van afwijkingen tussen gesynthetiseerde en werkelijke verdelingen. Door een Monte Carlo-steekproefprocedure worden meerdere synthetische populaties gegenereerd, waarna de beste op basis van chi-kwadraatstatistieken wordt geselecteerd.

Een belangrijk voordeel van IPU is het vermogen om problemen zoals lege cellen en nulmarges aan te pakken. Hiervoor worden aangepaste correcties toegepast, zoals het lenen van frequenties uit grotere geografische gebieden of het instellen van minimale drempelwaarden. De auteurs behandelen deze complicaties uitgebreid en laten zien dat deze aanpassingen essentieel zijn voor het toepassen van het algoritme in de praktijk.

De resultaten tonen aan dat IPU substantieel betere persoonsverdelingen oplevert dan traditionele IPF-methoden, terwijl de huishoudverdelingen exact worden gematcht. De methode is breed inzetbaar in transportmodellering en microsimulaties en biedt ook mogelijkheden voor het kalibreren van gewichten in mobiliteitsenquêtes en andere toepassingen waar meerdere verdelingen tegelijk moeten worden benaderd.

Analyse

1. Methode. Het artikel introduceert de *Iterative Proportional Updating (IPU)*-methode als een heuristisch algoritme dat zowel huishoud- als persoonskenmerken in synthetische populaties gelijktijdig probeert te matchen. Het algoritme wijzigt huishoudgewichten iteratief totdat de gemarginaliseerde distributies van beide niveaus overeenkomen met censusdata. De methode bouwt voort op Iterative Proportional Fitting (IPF), maar breidt deze uit met gewichtsaanpassingen die ook persoonsdistributies beïnvloeden.

2. Inputdata. De methode maakt gebruik van microdata uit de Amerikaanse *Public Use Microdata Sample (PUMS)* voor huishoud- en persoonskenmerken, gecombineerd met marginale distributies uit de *Census Summary Files* (voor controlevariabelen). Twee

blockgroepen uit Maricopa County, Arizona worden gebruikt als casus, met gegevens over onder meer inkomen, huishoudtype, leeftijd, geslacht en etniciteit.

3. Kenmerken. De gegenereerde populatie bestaat uit huishoudens en personen met gewichten die overeenkomen met de gemarginaliseerde verdelingen van beide niveaus. De methode ondersteunt meerdere categorische controlevariabelen en laat toe dat huishoudens binnen één type verschillende gewichten krijgen, afhankelijk van persoonskenmerken.

4. Schaal. De methode wordt toegepast op het niveau van *blockgroups*, de kleinste geografische eenheid waarvoor censusdata beschikbaar is. Er is een toepassing op 2.088 blockgroepen in Maricopa County met meer dan 3 miljoen inwoners, wat de schaalbaarheid aantoont.

5. Output. De output bestaat uit een gesynthetiseerde populatie met gewogen huishoudens en personen. Deze synthetische dataset is geschikt voor gebruik in activity-based of agent-based modellen. De gewichten worden gebruikt voor probabilistische steekproeftrekking, en de prestaties worden geëvalueerd met χ^2-statistieken en visuele vergelijkingen van joint distributions.

6. Validatie. Validatie vindt plaats op twee manieren: intern door vergelijking van de gegenereerde en oorspronkelijke distributies (bijv. χ^2-tests), en extern door visuele beoordeling van de fit met scatterplots. Meerdere steekproeven worden getrokken en de best passende wordt geselecteerd.

7. Openbaarheid. Hoewel de methode gedetailleerd beschreven wordt, is er geen openbare software of code gepubliceerd. De implementatie is gedaan in Python en MySQL, en het systeem is paralleliseerbaar, maar de code is niet beschikbaar gesteld.

8. Toepasbaarheid voor SIVMO. De IPU-methode is goed toepasbaar binnen het SIVMO-kader. De aanpak is geschikt voor kleine geografische eenheden, werkt met bestaande censusbronnen, en kan populaties genereren die compatibel zijn met agent-based simulatiemodellen. Beperkingen zijn dat het algoritme gevoelig is voor sparse data en dat categorisering zorgvuldig moet gebeuren om convergentie-problemen te vermijden.

Bijlage 3 Interview verslagen

3.1 PTV

PTV's population synthesiser
PTV recently integrated a population synthesiser directly into Visum, replacing its former dependence on the external PopulationSim tool. Until now, users had to install PopulationSim as a Python package and import its results, a process that often failed because of missing dependencies, restricted permissions, or opaque error messages. Many clients, particularly those working within large public organisations, were unable to run the software at all. What should have been a simple preparatory step for transport modelling had turned into a constant technical struggle.

The new built-in module, first released with Visum 2026, makes population synthesis part of the normal modelling workflow. It can now be run from within the software, without any external installation or separate configuration. PTV described this change as a pragmatic response rather than a strategic innovation. The logic remains close to PopulationSim, but the implementation is easier for non-programmers to use. Earlier attempts to help users through an installer or simplified wrapper had proved unreliable and were eventually abandoned.

Maintenance issues also played a role. PopulationSim, though conceptually strong, had not been updated properly. Some functions had stopped working, and dependency conflicts made the tool increasingly unstable. PTV had reported bugs and even improved installation, yet problems persisted. Over time, supporting the package became impossible. Creating an in-house alternative was therefore the only viable option.

A further reason lay in company policy. PTV has gradually reduced its reliance on external Python components to improve stability and security. Python remains available for scripting, but core functions are now maintained within the main codebase. The integrated synthesiser fits this approach: methodologically familiar, but more reliable.The goal was to develop "a system that works for everyone, not just the specialists".

Methodological approach
PTV's new synthesiser is built on the same conceptual foundation as PopulationSim. The method relies on entropy maximisation, also known as *list balancing*, to estimate weights for each household and person so that aggregated characteristics match the marginal control totals with minimal additional assumptions. Once these continuous weights are obtained, a linear-programming routine integerises the results, producing whole households and persons rather than fractional records. This combination of simultaneous balancing and integerisation ensures internal consistency across geographic levels and prevents the propagation of rounding errors.

The decision to retain this approach was straightforward. When PopulationSim was first introduced, PTV consulted academic experts who confirmed that entropy-based optimisation performed as well as other proportional fitting techniques if the input data were sound. Rather than investing in a new algorithm, PTV chose to preserve what already worked and focus on improving its usability and maintenance. The company views this as a continuation of the PopulationSim concept.

PTV also pointed out that most synthesis problems are not mathematical but practical. Inconsistent or incomplete data remain the main source of uncertainty. In Germany, official datasets such as national grids and survey aggregates often conflict once compared across scales. Correcting such discrepancies has a greater effect on model quality than choosing one optimisation method over another. For that reason, the company treats the algorithm as a reliable constant.

Applications and model types

The integrated synthesiser is used for both microscopic and macroscopic models. In agent- or activity-based frameworks, it generates populations of persons and households that form the basis for behavioural simulation. For traditional, trip-based models, the same process can run. This dual design allows one consistent synthesis step across different modelling traditions.

Within PTV, the synthesiser also feeds into the automated model-generation environment known as Model2Go. In that workflow, population creation forms the first link in a chain that automatically builds a full demand model. Having the synthesiser embedded in Visum simplifies the process and avoids external scripts or file conversions. The company expects that over time, population synthesis will become a standard component of every model development.

PTV views this integration as part of a wider modernisation of modelling practice. Many agencies still balance household and person categories manually in spreadsheets. Such methods are cumbersome and error-prone. By embedding a formal synthesiser into Visum, PTV hopes to standardise a transparent, reproducible process that can be reused for any region or scale. The result is both more efficient and more credible.

Data sources and requirements

In Germany, PTV relies mainly on the national household travel survey *Mobilität in Deutschland* (MiD) as its seed dataset. The survey includes more than 100 000 households and serves as the micro-sample from which synthetic populations are constructed. Each household is tagged by regional type, urban, semi-urban or rural, allowing the synthesiser to allocate appropriate households to each zone. Where survey coverage is sparse, PTV supplements it with aggregated control totals to maintain a plausible distribution.

The MiD survey is repeated roughly every five years, though the interval can vary. PTV considers this frequency adequate: household structures evolve slowly, and demographic relationships remain stable over time. For behavioural data such as trip chains or activity patterns, more recent information is preferable but not essential. Using surveys older than ten years would only be justified for calibration or historical comparison.

A persistent difficulty is inconsistency between datasets. Even statistics produced by the same authority can disagree once aggregated or disaggregated to different spatial levels. Such discrepancies constrain how precisely the synthesis can reproduce control totals. PTV stressed that these limitations lie in the data ecosystem rather than in the algorithm itself, and they apply to all modelling tools in the field.

Privacy and data access constraints

Privacy legislation strongly influences how population data can be used. In Germany, detailed geographic coordinates and rich socio-demographic information are held in separate databases that cannot be merged directly. Analysts must work on secure servers managed by the national statistical office, and outputs are screened before release. This ensures confidentiality but makes testing slow and cumbersome.

Moreover, small cell counts are suppressed for privacy reasons: categories with fewer than three observations are left blank. This rule creates inconsistencies between local and national totals, complicating validation. Robert noted that the situation in the Netherlands is comparable. Access to CBS microdata is restricted to a handful of accredited institutes, and most researchers must work on site. In practice, both systems make high-resolution validation legally possible but operationally difficult.

Experiences and use cases

At the time of the interview, the integrated synthesiser in Visum had been publicly available for only a week, so large-scale experience was limited. Earlier uses of PopulationSim had taken place in pilots or research projects rather than operational models. For national macroscopic applications, PTV still relies on non-integerised populations that can take several weeks to compute for the whole of Germany. These serve as reference runs for evaluating performance and scalability.

Despite its newness, PTV expects the in-built synthesiser to be adopted quickly. Consultants can now create synthetic populations entirely within Visum, without installing Python or managing dependencies. This simplicity removes the main obstacle that had prevented wider use of PopulationSim. The company's UK branch and other international teams plan to apply the new tool soon, which will provide further testing and validation opportunities.

Although specific projects were not discussed in detail, PTV anticipates that the new system will become part of routine model development. The shift from external to integrated synthesis is expected to save time, reduce errors and make advanced methods accessible to a much broader group of users.

Future perspectives and collaboration

PTV intends to improve the synthesiser gradually. Near-term updates will focus on runtime efficiency, better regional constraints and clearer diagnostics. Over the longer term, the company plans to match or exceed the functionality of PopulationSim while retaining the advantages of full integration in Visum. The tool will continue to be developed as a standard component rather than a separate product.

On the subject of openness, PTV takes a pragmatic view. PTV recognises the value of open-source transparency but doubts that such software can be maintained reliably without dedicated resources. Most users, they argued, care less about the visibility of

code than about the credibility of the results. PTV therefore prioritises well-documented methods, reproducibility and professional support. Collaboration with universities and public agencies remains welcome, particularly for validation studies or shared datasets, as long as it does not compromise system stability.

Closing reflections

PTV focuses on practicality and continuity. The move from PopulationSim to an integrated Visum module was not driven by theoretical innovation but by the need for a stable, user-friendly solution. PTV regards population synthesis as an essential yet routine element of modelling, a tool that should work reliably in the background rather than demand constant technical care. PTV's approach shows how implementation details can determine success more than algorithms do. Usability, data quality and maintainability are seen as the decisive factors. By embedding the synthesiser within Visum, PTV has transformed a fragile open-source process into a dependable professional function. It is a quiet but significant step towards making population synthesis a normal, reproducible part of transport modelling practice.

3.2 DfT (UK)

Role of DfT and purpose of the work

The UK Department for Transport (DfT) is responsible for the National Trip End Model (NTEM) at. This model forecasts trip generation for Great Britain - England, Wales and Scotland - based on population, demographics, car ownership and household structures. The model is aggregate in nature: it estimates how many trips are made but does not keep track of individual people or households. It therefore provides flows and totals but not disaggregate information on who makes which trips.

In developing a new version, DfT intends to change that. The aim is to keep track of individuals, allowing the model to produce a complete synthetic population alongside the aggregate trip forecasts. This would preserve compatibility with existing appraisal tools but also make the data suitable for activity-based or agent-based models that require detailed, person-level information. The goal is to create a hybrid system - one that continues to serve the existing modelling community while opening up opportunities for more advanced approaches.

The new synthetic population would be fully consistent with official assumptions about demographics, employment and households. It would form the foundation for trip generation forecasts and could also be published as a separate dataset. According to DfT, this would give the national models the same level of standardisation for activity-based work as the traditional four-stage models already enjoy.

Method and technical approach

The approach begins with microdata from the UK National Travel Survey (NTS). This dataset provides household and person-level samples that serve as the seed population. These records are scaled to national totals through a process of iterative sampling and adjustment. Growth factors are derived from Office for National Statistics (ONS) projections for population and households. The system therefore combines microdata sampling with official macro-level forecasts.

Under the hood, the process is executed using PopSim, a Python-based population synthesis engine. The software orchestrates multiple PopSim runs across regions, each constructed from ONS base data, growth factors and survey microdata. The results are then merged and re-aggregated to form a single national dataset containing households, persons and their characteristics. Cars are not modelled as individual entities, but household car ownership is included as one of the household dimensions.

The choice of PopSim was deliberate. At the start of the project, several alternatives were reviewed, including Spenser, a synthesiser originally developed for health-related applications. PopSim proved more flexible for modelling households and housing stock and offered a clear public-domain foundation on which to build. “Rolling your own from scratch,” DfT remarked, “is a mug’s game for wasting money and making amazing mistakes.” Adopting an existing, transparent platform allowed DfT to focus its resources on orchestration, validation and long-term maintainability rather than reinventing the algorithm.

Base year, data sources and treatment of COVID effects
The current development uses a 2021 base year, aligned with the most recent national census. DfT acknowledged that the data reflect COVID-era distortions, particularly in travel-to-work information, but concluded that they remain preferable to the outdated 2011 census. “Even if it’s bad, it’s still better than 2011, and there’s nothing better until 2031.” The 2021 dataset therefore provides the most recent and internally consistent demographic baseline available.

In addition to the census, the model uses the National Travel Survey as microdata input, omitting the pandemic years. Earlier and later waves are combined to provide a balanced sample unaffected by lockdown behaviour. Forecasts draw on external projections for population, households and employment, all sourced from ONS and other official agencies. Housing policy remains a difficult input: communication with the Ministry of Housing had been limited, leaving uncertainty over future housing distribution and capacity constraints. The software is therefore designed to accommodate alternative housing scenarios once these data become available.

Privacy and data handling
DfT seeks to avoid confidential data wherever possible. For business information, the team has switched from the confidential version of the business register to the public version, which provides coarser geographic detail but avoids legal restrictions. The trade-off in accuracy was judged acceptable given the administrative burden attached to confidential sources.

The NTS is used in a restricted but non-confidential form. Three versions of the survey exist: an unrestricted public file, a lightly restricted file available through agreement, and a secure version containing sensitive variables. The DfT project uses the intermediate version, which is sufficient for synthesis while remaining easy to access for licensed partners. DfT emphasised that the software pipeline is designed so that users can supply their own authorised NTS file without modification to the rest of the system.

Implementation and intended applications
The new system is being implemented as a complete software pipeline rather than a single synthesis run. It automatically downloads and prepares all required input data, generates PopSim control files, executes multiple runs in parallel for different parts of the country, and merges the outputs into a single integrated population. The pipeline also performs anonymisation, ensuring that individual records cannot be traced back to the original survey respondents.

In operation, the synthetic population will serve several functions. It will underpin the national trip generation forecasts for DfT’s official modelling guidance, providing a common demographic base for local and national transport models. It will also be usable by activity-based models, including a national ABM currently being explored by the department, allowing such models to adopt the same underlying assumptions as the strategic ones.

Beyond transport, DfT sees potential applications in survey analysis and digital twins. Local synthetic populations could be used to improve weighting procedures in survey samples or to explore policy impacts at fine geographic scales. The detailed household

and person records, each linked to Middle Layer Super Output Areas (MSOAs), make it possible to integrate the data into simulation environments or microsimulation frameworks beyond transport.

Collaboration, quality assurance and publication
The project is delivered by Arup under contract to DfT, working with academic partners including Nik Lomax and colleagues from the University of Leeds and University College London. DfT designed the tender to encourage such collaboration, ensuring both technical expertise and independent peer review. DfT monitors progress through GitHub, where all code changes and issues are logged, enabling full version control and transparency.

Quality assurance combines internal and external scrutiny. Arup performs code reviews, calibration and formal documentation, while DfT conducts acceptance tests and reasonableness checks once the software is handed over. DfT expects that publication will itself act as an additional layer of validation: “If we publish, people will scrutinise the hell out of it. If we don’t, they’ll reverse-engineer it and accuse us anyway.” Public release of both data and methods is therefore viewed as the most effective defence against errors and criticism.

Publication is a central ambition. The plan is to publish not only the aggregate trip-end results but also the synthetic population itself and, if possible, the orchestration software that links data to PopSim. Open-sourcing the software would allow local authorities, consultants and academics to create their own scenarios, provided they label them as unofficial. DfT acknowledged that this would require approval from senior management but believes it aligns with DfT’s open-data policy and would strengthen transparency across the modelling community.

Challenges and maintenance
The greatest challenge, DfT noted, will not be technical but institutional. Maintaining software of this complexity demands constant updates and support. Dependencies evolve, components age and security standards change. Keeping the system running requires continuous attention - “You have to run just to stand still.” Another concern is misuse: once the results are public, users may treat them as more certain than they are. Even detailed documentation and caveats often go unread. To counter this, DfT plans ongoing guidance and communication, ensuring that users understand the scope and limitations of the data.

Political and legal considerations add further complexity. The model must adhere to government standards for intellectual property, accessibility and quality assurance while remaining flexible enough for external collaboration. DfT sees maintaining that balance - between control and openness - as the main long-term test of success.

Advice for future development in the Netherlands
Asked what advice he would give to organisations developing population synthesis in the Netherlands, DfT offered a mixture of practical and strategic guidance. They urged teams to build on existing standards wherever possible, whether technical (such as coding or documentation standards) or procedural (such as ISO quality frameworks). Equally important, he said, is transparency: publishing assumptions, methods and

intermediate results helps to build trust and allows others to contribute constructively.

DfT also emphasised the value of community and continuity. Collaboration across agencies, academics and consultants ensures that expertise is shared and knowledge survives beyond individual projects or personnel changes. Open-source development, if properly curated, can help maintain an ecosystem of users and contributors that keeps the work alive even when official teams move on.

Finally, DfT advised starting with a clear, manageable set of attributes and expanding only when data and confidence allow. Models that try to serve too many domains or scales risk collapsing under their own weight. Starting simple and building robustness first, he argued, is the surest way to long-term success.

Closing reflections

The interview with DfT revealed a pragmatic and candid view of population synthesis within government. The DfT's initiative aims to modernise its national modelling framework by moving from aggregate trip generation to a person-level synthetic population, combining rigour with accessibility. The project relies on open methods, tested software and academic collaboration rather than bespoke development.

DfT's emphasis on transparency, maintenance and clear communication highlights how national data infrastructure depends as much on governance as on mathematics. The ambition - to publish the synthetic population and the software pipeline itself - illustrates a shift towards openness and reproducibility within the UK's modelling community. If realised, it could give both official and independent modellers a common foundation for future appraisal and analysis, setting a useful precedent for other countries pursuing similar goals.

3.3 TNO

Achtergrond en aanleiding
TNO is al geruime tijd actief op het gebied van populatiesynthese. De eerste stappen werden ongeveer acht jaar geleden gezet, toen het instituut begon met de ontwikkeling van een eigen populatiegenerator. De aanleiding was de opzet van een nieuw activity-based model binnen het project Urban Tools Next. Onder leiding van Erik de Romph koos TNO destijds voor de Iterative Proportional Fitting (IPF)-methode, een gangbare en relatief overzichtelijke techniek binnen de wereld van verkeers- en vervoersmodellen. Als belangrijkste databron diende de CBS-microdata. De initiële implementatie werd in MATLAB gerealiseerd en in de loop der jaren uitgebreid met extra variabelen en validaties.

De eerste toepassing richtte zich op de regio Rotterdam–Den Haag (MRDH), maar inmiddels is er een landelijke populatie beschikbaar op postcode-4-niveau. Deze populatie wordt gebruikt binnen diverse activity-based modellen. De dataset is recent nog geactualiseerd op basis van CBS-data uit 2023, met aanvullende schattingen voor toekomstjaren, onder andere voor 2030.

Methode en implementatie
De gekozen IPF-methode werkt met drie basisvariabelen: leeftijd, geslacht, herkomst en een extra wisselende variabele zoals inkomen, autobezit, opleidingsniveau, rijbewijs of huishoudtype. Op die manier ontstaan meerdere deelpopulaties die elk afzonderlijk worden berekend. TNO probeerde aanvankelijk meerdere dimensies tegelijk te verwerken, maar dat bleek instabiel. Daarom wordt elke extra dimensie onafhankelijk aan de drie basisvariabelen toegevoegd.

Een bekend nadeel van deze aanpak is dat de onderlinge relaties tussen de ‘derde variabelen’ verloren kunnen gaan. De marges kloppen, maar de afhankelijkheden, bijvoorbeeld tussen inkomen en huishoudgrootte, worden niet volledig behouden. Om dit te beperken voert TNO enkele postprocessing-stappen uit. Zo wordt achteraf gecontroleerd of personen zonder rijbewijs geen auto’s bezitten en wordt autobezit binnen huishoudens aangepast op basis van de aanwezige rijbewijsbezitters.

De koppeling tussen individuen en huishoudens verloopt via een heuristische, iteratieve procedure. Daarbij wordt herhaaldelijk gecontroleerd of het aantal huishoudens van elk type overeenkomt met de verdeling in het gebied. Ouders en kinderen worden bijvoorbeeld realistisch aan elkaar gekoppeld, rekening houdend met leeftijdsverschillen. Voor de afronding naar gehele personen gebruikt TNO een iteratief algoritme dat afwijkingen minimaliseert, vergelijkbaar met het toekennen van restzetels in verkiezingssystemen.

Databronnen en privacy
De synthese is gebaseerd op de CBS-microdata, die binnen de beveiligde CBS-omgeving worden gebruikt. De populatie wordt op postcode-4-niveau (PC4) geconstrueerd, omdat fijnere niveaus, zoals PC5 of PC6, te privacygevoelig zijn en daardoor niet geëxporteerd mogen worden. Volgens CBS-regels mogen categorieën pas worden vrijgegeven wanneer er minstens tien personen aan een combinatie van

kenmerken voldoen. Voor dichtbevolkte gebieden lukt dat soms op PC5-niveau, maar voor landelijke dekking is PC4 het hoogst haalbare.

De gebruikte microdata omvatten onder meer bestanden over inkomens (Spolis), bedrijven (ABR), rijbewijzen, huishoudens en demografie. Waar categorieën te klein zijn om direct te gebruiken, worden waarden geïmputeerd. Dat gebeurt door schattingen te maken die ervoor zorgen dat het totaal per zone consistent blijft. Personen worden vervolgens willekeurig aan gebouwen gekoppeld binnen hun postcodegebied. Zo blijft de koppeling tussen individuen en woningen realistisch maar anoniem.

Validatie en nauwkeurigheid
De belangrijkste validatie vindt plaats op zoneniveau. Daarbij wordt gecontroleerd of de marges overeenkomen met de CBS-verdelingen op PC4-niveau. Daarnaast voert TNO landelijke validaties uit, waarbij correlaties tussen variabelen op nationaal niveau worden vergeleken met de werkelijkheid. In 2026 wil men binnen de CBS-omgeving ook validaties op persoonsniveau uitvoeren, om de interne consistentie beter te kunnen toetsen. Kleine afwijkingen blijven aanwezig, maar over het algemeen sluit de synthetische populatie goed aan bij de officiële statistieken.

Toepassingen
De gegenereerde populatie wordt gebruikt binnen twee hoofdmodellen: het activity-based model van TNO, dat draait op de open-source software ActivitySim, en de New Mobility Modeler, een gedesaggregeerd vervoerwijzekeuzemodel. In beide gevallen vormt de populatie de input voor simulaties van verplaatsingsgedrag. Daarnaast wordt de data ingezet voor beleidsanalyses en indicatorontwikkeling, bijvoorbeeld op het gebied van ruimtegebruik, transportarmoede, energiearmoede en gezonde leefstijl. Door de populatie te koppelen aan andere databronnen ontstaan nieuwe mogelijkheden om sociaal-demografische en ruimtelijke patronen in samenhang te analyseren.

De toepassingen variëren van wetenschappelijk onderzoek tot opdrachten voor gemeenten, provincies en nationale overheden. Zo is de populatie gebruikt in het MRDH-model en binnen het NWO-project XCARCITY, dat de relatie tussen ruimtelijke schaarste en mobiliteitsgedrag onderzoekt.

Ervaringen en uitdagingen
Volgens TNO functioneert de methode goed voor de meeste toepassingen, maar het voorbereiden van de inputdata blijft tijdrovend en arbeidsintensief. Vooral de bewerking van toekomstscenario's vergt veel handwerk. Voor de toekomstjaren (bijvoorbeeld 2030) worden aannames gebruikt over de verdeling van kenmerken, vaak op basis van landelijke modellen zoals het LMS. Wanneer geen specifieke toekomstdata beschikbaar zijn, worden relaties uit het meest recente basisjaar overgenomen en toegepast op nieuwe marges.

Een complicatie is de koppeling van de landelijke populatie (PC4) aan regionale verkeersmodellen met een eigen zonering, zoals het MRDH- of VMA-model. Voor dergelijke toepassingen wordt de populatie soms gedesaggregeerd naar kleinere zones, op basis van BAG-gegevens over gebouwen. De kwaliteit van deze verdelingen is niet altijd bekend en kan variëren per gebied. Tegelijkertijd is detailniveau belangrijk

voor stedelijke analyses, omdat buurten sterk kunnen verschillen in sociaaleconomisch profiel en mobiliteitsgedrag.

Voor toekomstscenario's vormt het ontbreken van een transitiemodel een beperking. Veranderingen in populatiesamenstelling, migratie en woningbouw worden nu slechts impliciet meegenomen via aangepaste marges. Een meer dynamische aanpak, vergelijkbaar met methoden in TigrisXL waarin bevolkingsontwikkeling per jaar wordt voorspeld, zou nauwkeurigere resultaten opleveren, maar vraagt om extra data en modellering.

Organisatie en samenwerking
TNO ziet het als een positief initiatief om te werken aan een landelijke populatie-synthese die bruikbaar is voor verschillende overheden en modellen. Daarbij is het volgens hen logisch om de CBS-microdata als basis te gebruiken, vanwege de rijkdom en detailgraad van deze bron. Het nadeel is dat deze data alleen toegankelijk zijn voor onderzoeksinstellingen en universiteiten, waardoor marktpartijen uitgesloten blijven.

Een mogelijke oplossing is een samenwerkingsconstructie waarbij onderzoeks-instituten met CBS-toegang periodiek geaggregeerde bestanden produceren die vervolgens breder gedeeld kunnen worden. TNO acht dit haalbaar, mits de juridische voorwaarden met het CBS duidelijk worden vastgelegd. Een vergelijkbare werkwijze bestaat al voor het thema energiearmoede, waarbij TNO en CBS gezamenlijk analyses uitvoeren en resultaten periodiek publiceren.

Over samenwerking met andere partijen is TNO positief, mits er voldoende ruimte blijft voor eigen onderzoek en innovatie. De onderzoekers verwijzen naar de RWS-SMP-alliantie als een werkend voorbeeld van een gesloten maar samenwerkende community. Een dergelijk model, waarin kennisinstellingen, consultants en overheden bijdragen aan één gedeelde tool, zou volgens TNO ook voor populatiesynthese goed kunnen functioneren.

Aanbevelingen
Een belangrijke aanbeveling van TNO is om in toekomstige populaties expliciet rekening te houden met afwijkende huishoudtypen zoals studentenhuizen, bejaarden-tehuizen en gevangenissen. Deze groepen wijken sterk af van het gemiddelde en verdienen daarom een aparte behandeling in de synthese. Daarnaast wijst TNO op het probleem van leaseauto's, die in de data vaak aan bedrijven zijn gekoppeld in plaats van aan huishoudens. Dit leidt tot onnauwkeurigheden in autobezitstatistieken. Voor een betere toewijzing verwijzen zij naar onderzoek van het KIM in het kader van de Autoatlas, dat hier bruikbare inzichten kan bieden.

Slotbeschouwing
Het gesprek met TNO laat zien dat het instituut beschikt over een volwassen, goed onderhouden populatiesynthese die breed toepasbaar is binnen en buiten het vervoersdomein. De methode is pragmatisch en robuust, maar tegelijk beperkt in het aantal dimensies en de mate waarin relaties tussen variabelen behouden blijven. De grootste uitdagingen liggen niet in de algoritmen, maar in data, validatie en organisatie: toegang tot microdata, consistentie van bronnen en de afstemming tussen partijen.

TNO pleit voor een gezamenlijke aanpak waarin onderzoeksinstellingen en markt-partijen samenwerken aan een landelijke, open toepasbare populatie, met periodieke updates en heldere afspraken met het CBS. Daarmee kan populatiesynthese uit-groeien tot een stabiele bouwsteen voor tal van beleidsdomeinen – van verkeer en vervoer tot energie, ruimte en gezondheid.

3.4 CBS

Context en achtergrond van het gesprek
De aanleiding voor het gesprek was de toenemende behoefte binnen SIVMO en de Nederlandse transportmodelpraktijk om populatiesynthese structureel en verantwoord mogelijk te maken. Voor veel modellen, zeker activity-based of agent-based modellen, zijn microdata noodzakelijk, maar de toegang tot deze bronnen is juridisch en technisch sterk begrensd.

CBS gaf aan dat dit gesprek plaatsvindt in een periode waarin het thema *synthetische data* steeds meer aandacht krijgt, zowel beleidsmatig als wetenschappelijk. Tegelijkertijd staat het onderwerp maatschappelijk onder druk door politieke vragen en zorgen over privacy. Hierdoor beweegt CBS in een spanningsveld tussen innovatie en wettelijke beperkingen. Het gesprek bood inzicht in hoe CBS deze balans ziet, welke mogelijkheden er zijn voor populatiesynthese en waar de grenzen liggen.

Historische ervaring van CBS met populatiesynthese
CBS heeft in het verleden geen eigen populatiesynthesetool ontwikkeld, maar wél actief bijgedragen aan projecten waarin populatiesynthese een centrale rol speelde. Een vroeg voorbeeld is het traject rond FEATHERS in de regio Rotterdam, ongeveer zeven à acht jaar geleden. In dat project werden CBS-microdata gebruikt in de veilige omgeving, waarna een externe partij (TNO) een synthetische populatie genereerde. CBS hield toezicht op de databeveiliging en voerde outputcontroles uit, maar had geen actieve betrokkenheid bij de methodologische keuzes.

Volgens de gesprekspartners komt populatiesynthese binnen CBS meestal voor onder de noemer “synthetische data”. Dat is een bredere categorie waarin ook ruis geïntroduceerde datasets en statistisch gegenereerde varianten van microdata vallen. CBS ziet een toenemende vraag naar dit soort toepassingen, maar ook een toename van risico’s. De organisatie is daarom terughoudend in het toestaan om resultaten van complexe bewerkingen die mogelijk privacyrisico’s met zich meebrengen naar buiten de RA-omgeving te brengen.

Daarnaast is CBS betrokken in diverse trajecten waarin synthetische of afgeleide datasets worden ontwikkeld voor specifieke beleidsdomeinen, zoals energiearmoede of gezondheidsmonitoring. Deze projecten laten zien dat synthetische data waardevol kunnen zijn, maar CBS benadrukt dat ze nooit gelijk staan aan microdata en dat transparantie over de beperkingen essentieel is.

Data, toegankelijkheid en detailniveaus
CBS gaf een uitgebreid overzicht van de beschikbare databronnen en toegangsvoorwaarden. Toegang tot microdata verloopt via de Remote Access (RA)-omgeving, een beveiligde werkomgeving waarin erkende onderzoeksinstellingen op afstand met gevoelige data kunnen werken. De data blijven binnen CBS, en gebruikers mogen alleen geaggregeerde, gecontroleerde output exporteren.

Voor microdata gelden strikte regels:
- Alleen onderzoekers van instellingen met een RA-machtiging kunnen toegang krijgen.

- Projecten moeten worden goedgekeurd op basis van doelbinding, proportionaliteit en privacybeleid.
- Marktpartijen kunnen ook toegang krijgen to de RA-omgeving mist ze de procedure om een machtiging te krijgen succesvol hebben doorlopen.

Het detailniveau dat in RA beschikbaar is, is vaak postcode-6 (PC6), maar buiten RA geldt een sterke beperking. Sinds 2024 mogen semi-open datasets (zoals die via DANS) niet verder gaan dan postcode-4 (PC4). Bovendien zijn deze bestanden verrijkt met ruis via de PRAM-techniek (Post Randomization Method), waardoor exacte waarden zijn vervaagd en sommige variabelen in categorieën zijn opgedeeld. CBS gaf aan dat deze maatregelen noodzakelijk zijn vanwege toenemende risico's op herleidbaarheid en steeds geavanceerdere analysetechnieken. De adviseurs van de CBS Microdata Services zijn altijd bereid om meer gedetailleerde en specifieke informatie te verschaffen- RA-aanvragen zijn maatwerk.

Voor populatiesynthese betekent dit dat hoogwaardige brondata in beginsel alleen binnen RA kunnen worden gebruikt. Het genereren van een volledige populatie buiten de omgeving is volgens CBS lastig als de synthese microdata nodig heeft op persoonsniveau en huishoudniveau. De organisatie is duidelijk: populatiesynthese met echte microdata kan alleen in RA plaatsvinden, tenzij duidelijk is aangetoond dat de privacy-risico's afdoende zijn afgedekt.

Privacy, juridische kaders en risico-inschatting
CBS werkt binnen een strikt juridisch kader, gebaseerd op de CBS-wet, de AVG en Europese richtlijnen. De functionaris gegevensbescherming hanteert het principe dat synthetische data in eerste instantie worden beschouwd als een bewerking van echte data, en dus onder dezelfde regels vallen.

Het CBS maakt onderscheid tussen:
- Directe identificaties (geslacht, leeftijd, postcode),
- Indirecte identificaties (combinaties van variabelen),
- Herleidbaarheid via patronen (met name door moderne AI of machine-learningmodellen).

De risico-inschatting is de afgelopen jaren scherper geworden. CBS wees expliciet op de technologische trends waarin machine-learningmodellen in staat zijn om synthetische data terug te redeneren naar oorspronkelijke individuen. Deze zogeheten *model inversion attacks* en *membership inference attacks* worden door ethici en juristen gezien als realistische risico's.

CBS signaleerde ook een maatschappelijke gevoeligheid rond het gebruik van synthetische data. Er werd verwezen naar een rapport van het Tilburg Institute for Law and Technology (TILT), getiteld *Op weg naar een synthetische samenleving,* waarin zorgen werden geuit over de mogelijke inzet van synthetische data, vooral in de vorm van deep fakes. Dat kan afstralen op gebruik in de publieke sector en vergroot de kans dat burgers het vertrouwen in statistische instituties verliezen. Parlementaire vragen over dit onderwerp geven aan dat de risico's reëel zijn. Deze en andere observaties hebben bijgedragen aan de huidige voorzichtige beleidslijn binnen CBS.

Kwaliteit, validatie en betrouwbaarheid
CBS benoemde dat er op dit moment geen breed geaccepteerde standaard bestaat voor het valideren van synthetische populaties op aspecten als inhoudelijke bruikbaarheid ('fit for use') en onthullingsrisico's. Projecten zoals FEATHERS laten zien dat synthetische datasets nuttig zijn, maar het blijft lastig vast te stellen welke variabelen goed zijn gesynthetiseerd en waar onnauwkeurigheden optreden.

Om die reden werkt CBS samen met onder meer TNO, DUO, Belastingdienst, UWV en verschillende universiteiten aan een project om richtlijnen te ontwikkelen voor kwaliteitsborging van synthetische gegevens. Een belangrijk doel is het ontwikkelen van een beoordelingskader dat zowel de statistische kwaliteit als de privacybescherming meeneemt.

CBS verwacht dat populatiesynthese een blijvend onderdeel wordt van onderzoek, maar dan wel onder duidelijke spelregels:
- transparantie over aannames,
- expliciete waarschuwingen bij interpretatie,
- duidelijke communicatie over beperkingen en onzekerheden.

Toepassing en gebruik door externe partijen
CBS ziet een groeiende markt waarin partijen zoals Syntho en BlueGen actief synthetische data genereren. Beide bedrijven beschikken over expertise in machine learning en simulatie en het CBS werkt met beide bedrijven samen, maar CBS blijft kritisch. Er bestaan zorgen over transparantie: niet alle partijen delen hun methoden en validaties, wat het risico vergroot dat synthetische data beleidsmatig worden gebruikt zonder dat de herkomst of kwaliteit goed te beoordelen is. Dat kan leiden tot verkeerde conclusies die beleidsmatig doorwerken.

Daarom werkt CBS slechts onder strikte voorwaarden met dit soort bedrijven samen. Met BlueGen loopt een experimenteel traject met gekoppelde datasets waarin CBS toetst hoe veilig synthetische data zijn en of deze datasets kunnen worden vrijgegeven zonder privacyrisico's. Hoewel de voorlopige resultaten bemoedigend zijn, benadrukte CBS dat het nog te vroeg is om synthetische data als volwaardig alternatief voor microdata te beschouwen.

Organisatie, toegang en samenwerking met SIVMO
Een belangrijke vraag tijdens het gesprek was of een collectief zoals SIVMO toegang kan krijgen tot microdata. CBS gaf aan dat dit hoogstwaaarschijnlijk niet mogelijk is: SIVMO is geen onderzoeksinstelling, heeft geen wettelijke onderzoeksstatus en kent geen eigen rechtspersoonlijkheid. Alleen instellingen met formele machtiging, universiteiten, planbureaus, sommige onderzoeksinstituten zoals TNO, kunnen toegang krijgen.

Wel is er ruimte voor samenwerking (hoe dit er precies uit moet komen te zien vraagt een nadere uitwerking, inclusief juridische toetsing):
- Een gemachtigde instelling kan namens SIVMO in RA opereren.
- Synthese kan dan door die instelling worden uitgevoerd, onder toezicht van CBS.
- Resultaten kunnen vervolgens worden geaggregeerd en gecontroleerd geëxporteerd.

Een model waarin één of meerdere CBS-gemachtigden periodiek een landelijke synthetische populatie genereren voor alle SIVMO-partners ziet CBS als een werkbare optie, mits privacy en governance goed geregeld zijn.

Toekomst en mogelijke oplossingsrichtingen
CBS verkent momenteel technologische mogelijkheden om het werken met microdata te vergemakkelijken zonder dat privacy in gevaar komt. Een van de belangrijkste opties is het gebruik van SURF High Performance Computing (HPC) via OSSC. Daar zou het in principe mogelijk zijn om modellen extern te ontwikkelen en deze vervolgens binnen een besloten CBS-omgeving te laten draaien. Dat vereist een streng scheiding tussen ontwikkelomgeving en executieomgeving, maar zou de kloof tussen microdata en modellen verkleinen.

Daarnaast speelt ODISSEI een belangrijke rol. Binnen dat platform wordt gewerkt aan bredere toegankelijkheid van onderzoeksdata, inclusief synthetische varianten. CBS ziet kansen om populatiesynthese te verbinden met dit ecosysteem, vooral waar sociaal-demografische datasets relevant zijn voor mobiliteit, gezondheidszorg of energieonderzoek.

Slotbeschouwing
Het gesprek met CBS maakte duidelijk dat populatiesynthese in Nederland (en Europa) onlosmakelijk verbonden is met vraagstukken rond privacy, governance en maatschappelijk vertrouwen. CBS staat open voor samenwerking, maar uitsluitend binnen wettelijke en ethische kaders. De organisatie benadrukt dat kwalitatief goede synthetische populaties alleen kunnen worden ontwikkeld wanneer data verantwoord worden gebruikt, privacy aantoonbaar is beschermd en de beperkingen helder worden gecommuniceerd.

CBS ziet de waarde van populatiesynthese voor verkeers- en vervoersmodellen, maar waarschuwt voor het te snel aannemen dat synthetische data geen risico's kennen. De combinatie van technologische ontwikkelingen, maatschappelijke gevoeligheid en juridische verplichtingen maakt het noodzakelijk om voorzichtig en transparant te opereren. Voor SIVMO betekent dit dat een duurzame oplossing waarschijnlijk ligt in samenwerking met gemachtigde onderzoeksinstellingen en in het ontwikkelen van een gedeelde, gecontroleerde aanpak voor de Nederlandse modelpraktijk.

3.5 Goudappel

Families van populatiesynthese-methoden
Goudappel onderschrijft de indeling in zes hoofdgroepen die in het SIVMO-rapport is opgenomen, maar plaatst daar enkele kanttekeningen bij.

Volgens Goudappel is het onderscheid tussen de twee hoofdcomponenten van populatiesynthese, enerzijds *fitting* (het laten aansluiten op marges) en anderzijds *allocation* (het verdelen op individueel niveau), fundamenteel voor het begrijpen en vergelijken van methoden. Fitting is typisch van belang voor modellen die op marges vertrouwen, terwijl allocation cruciaal is voor microsimulaties en agent-based modellen, waarin individuen of huishoudens unieke combinaties van kenmerken moeten hebben.

Een indeling op basis van wiskundige techniek of algoritme (zoals IPF, IPU, entropie-optimalisatie, sampling of simulatie) is volgens hem niet voldoende. Goudappel wijst erop dat veel methoden in de praktijk een combinatie zijn van technieken die zowel fitting als allocation omvatten, of die in twee stappen gescheiden worden toegepast.

Daarnaast benadrukt Goudappel dat de meeste overzichtsliteratuur een platte lijst van technieken presenteert, terwijl het beter is om de methoden hiërarchisch of modulair te ordenen: welke onderdelen worden opgelost, in welke volgorde, en met welke techniek? Goudappel verwijst hierbij expliciet naar het werk van Müller & Axhausen (2010) en naar een afstudeerscriptie van De Jong waarin populatiesynthese wordt benaderd als een keten van stappen in plaats van als één methode. Deze aanpak maakt het ook mogelijk om combinaties van technieken of hybride oplossingen te beschrijven.

Goudappel stelt voor dat het eindrapport van SIVMO deze tweedeling, fitting en allocation, duidelijker naar voren brengt. Sommige tools gebruiken namelijk een eenvoudige fittingmethode maar combineren dat met een geavanceerde allocatiemodule (zoals Octavius), terwijl andere juist complexere technieken gebruiken om beide tegelijk te doen.

Tot slot werd besproken dat de opkomst van deep learning en generatieve modellen zoals VAEs of GANs een apart domein vormt. Deze methoden combineren impliciete correlatieherkenning met steekproefgeneratie, maar zijn minder transparant en minder breed inzetbaar in beleidspraktijk. Hun plaats in de indeling is nog in ontwikkeling.

Ontwikkeling en werking van Octavius
Octavius is ontwikkeld door Goudappel/Dat.mobility als generieke module voor populatiesynthese. De tool is afgeleid van de ‘Population Synthesizer’ die eerder als zelfstandige Java-module functioneerde. Gedurende de ontwikkeling is deze synthesizer aangepast, uitgebreid en geoptimaliseerd op basis van gebruikservaring, eisen van opdrachtgevers, en beperkingen van bestaande technieken.

De oorspronkelijke versie van de Population Synthesizer maakte gebruik van een standaard IPF-methode op persoonsniveau, waarmee persoonskenmerken op basis van marges werden gesynthetiseerd. Al snel werd een tweede IPF-stap toegevoegd

om ook huishoudkenmerken te modelleren. In de praktijk bleek dit tweestaps-IPF-model onvoldoende robuust in situaties waar sterke afhankelijkheden bestonden tussen huishoudens en personen (bijv. gezinssamenstelling en arbeidsparticipatie).

Om die reden werd geëxperimenteerd met IPU (Iterative Proportional Updating), dat zulke afhankelijkheden explicieter kan modelleren. De implementatie leverde echter instabiliteit op in sommige gevallen en beperkte controle over de integerisatie. Uiteindelijk is daarom gekozen voor een alternatieve benadering op basis van non-negative least squares (NNLS). Deze methode maakt het mogelijk om huishoudens en personen tegelijkertijd te synthetiseren, met behoud van de consistentie tussen aggregatieniveaus.

Een belangrijk onderdeel van Octavius is het discretisatie-algoritme dat gebruikt wordt om populaties met 'hele eenheden' te genereren. Hiervoor ontwikkelde Goudappel een techniek, aangeduid als SNET, die zorgt voor een zo klein mogelijke fout tussen de continue oplossing van het optimalisatieprobleem en de uiteindelijke discrete populatie. Dit in tegenstelling tot Monte Carlo-benaderingen waarbij steekproeven ruis introduceren.

De integerisatie vindt plaats ná de fittingstap, waarbij huishoudens en personen uit een pool worden geselecteerd. Deze selectie wordt gestuurd op basis van gewichten uit het optimalisatie-algoritme, zodat het eindresultaat voldoet aan alle marges binnen acceptabele foutmarges.

Octavius is modulair opgezet. De synthesizer is een zelfstandige module die losstaat van het verkeersmodel waarin het wordt ingezet. Het kan worden gekoppeld aan andere platformen Omnitrans. Dankzij deze modulaire structuur is het mogelijk om dezelfde tool te gebruiken voor verschillende doelen: scenario-opbouw, populatieactualisatie, huishoudsimulatie of synthetische steekproeftrekking.
De tool is geoptimaliseerd voor middelgrote datasets en categorische verdelingen. Octavius is in de praktijk voornamelijk gebruikt op het niveau van buurten, gemeenten of COROP-gebieden, maar ondersteunt ook landelijke toepassingen.

Data en randvoorwaarden
Octavius en de onderliggende Population Synthesizer zijn niet gebonden aan specifieke databronnen. De tools zijn ontworpen als generieke engines die gevuld kunnen worden met uiteenlopende combinaties van microdata en randtotalen. Deze modulariteit maakt het mogelijk om het systeem aan te passen aan de beschikbare bronnen per project of regio.

Voor fitting worden doorgaans CBS-buurtstatistieken, ODiN en MPN gebruikt. In sommige gevallen wordt ook gebruik gemaakt van registerdata, mits deze zijn geaggregeerd tot het vereiste schaalniveau. Voor specifieke kenmerken zoals voertuigbezit, arbeidsparticipatie of opleidingsniveau worden aanvullende marginale totalen berekend of afgeleid uit secundaire bronnen.

Een terugkerend probleem is dat veel databronnen niet alle benodigde variabelen bevatten of geen consistente definities hanteren. Zo bevat ODIN persoonskenmerken maar geen huishoudstructuur, terwijl MPN huishoudinformatie biedt maar op een te kleine steekproefgrootte voor landelijke toepassing.

Om dit op te lossen wordt gewerkt met zogenaamde imputatieregels: afhankelijkheden tussen kenmerken worden gereconstrueerd op basis van microdata uit andere bronnen, zoals het CBS microdatabestand (indien beschikbaar binnen projectcontext) of oudere databronnen zoals OVG of MON. Goudappel benadrukt dat de huishoudstructuur in Nederland relatief stabiel is, wat het mogelijk maakt om ook oudere datasets in te zetten, zolang deze voldoende volume hebben.

Gebruik van CBS microdata is juridisch en praktisch lastig. De bestanden mogen niet gedownload worden, en er zijn geen mogelijkheden om synthesemodules op de CBS-server te draaien. Voor populatiesynthese betekent dit in de praktijk dat er buiten-CBS-om gewerkt moet worden, wat leidt tot beperkingen in detailniveau en validatiemogelijkheden.

Voor het vullen van ontbrekende categorieën (zoals OV-kaartbezit) wordt soms teruggevallen op externe bronnen of expertinschattingen. In het geval van populatiesynthese over meerdere jaren (bijv. t.b.v. backcast of forecast) is er doorgaans sprake van losstaande syntheses per jaar, tenzij specifieke scenario's jaar-op-jaar coherentie vereisen.

De populatiesynthese kan gestuurd worden door randvoorwaarden zoals:
- vaste marges op huishoud- en persoonskenmerken (enkelvoudig en kruistabellen),
- uitsluitingen (bijv. onmogelijke combinaties),
- continue gewichten voor scenariofactoren (bijv. voertuigbezit of arbeidsparticipatie).

Octavius ondersteunt dit door een constraint-systeem waarin zowel harde als zachte voorwaarden ingevoerd kunnen worden. Dit maakt het model flexibel in toepassing, maar vraagt wel expertise bij het configureren van de randvoorwaarden.

Validatie en kwaliteit
Validatie van gesynthetiseerde populaties is belangrijk om vertrouwen te wekken in het gebruik ervan binnen verkeers- en vervoersmodellen. Tijdens het gesprek kwamen verschillende vormen van validatie aan bod, evenals de beperkingen die in de praktijk worden ervaren.

De meest basale vorm van validatie bij Octavius is de interne controle op marges. Dat wil zeggen: controleren of de synthetische populatie voldoet aan de ingevoerde marginale verdelingen, zowel op huishoudniveau als persoonsniveau. Deze validatie is in feite ingebouwd in de synthese zelf, omdat de methoden zijn ontworpen om precies aan deze marges te voldoen (bij methoden zoals NNLS of IPU).

Voor de meeste opdrachtgevers is margevalidatie voldoende: als de verdelingen per kenmerk kloppen op het gewenste schaalniveau (bijvoorbeeld gemeente of regio), wordt aangenomen dat de populatie bruikbaar is voor verdere modellering.

Een diepere vorm van validatie betreft het toetsen van de gegenereerde microdata aan onafhankelijke observaties. Dat kan bijvoorbeeld door:
- synthetische kruistabellen te vergelijken met tabellen uit CBS microdata;
- gedragseffecten (zoals voertuigbezit per huishoudenstype) te vergelijken met gegevens uit OVG of andere gedragsdatabronnen;

- output van het model te vergelijken met externe statistieken (bijv. verdeling OV-chipkaartbezit of mobiliteitsstatistieken).

In de praktijk is deze vorm van validatie minder gangbaar. De reden is tweeledig: enerzijds zijn de benodigde microdata niet beschikbaar of mogen ze niet worden gebruikt, anderzijds is er geen opdrachtgever die dit expliciet vraagt of financiert. Dat leidt ertoe dat deze validatiestap vaak overslagen wordt, ondanks de methodologische relevantie.

Een belangrijk kwaliteitsaspect is de stabiliteit van de uitkomsten. Octavius is ontworpen als deterministisch model met minimale afhankelijkheid van toeval. Dit betekent dat de uitkomsten bij gelijke invoer altijd gelijk zijn, een voordeel ten opzichte van Monte Carlo-gebaseerde synthesemethoden.

De integerisatie via SNET garandeert bovendien dat discretisatiefouten klein zijn en consistent blijven over runs. Daardoor is het model geschikt voor toepassing in contexten waar reproduceerbaarheid en betrouwbaarheid belangrijk zijn, zoals beleidssimulaties of modelprognoses over meerdere jaren.

Toch geldt dat bij complexe kruistabellen of lege cel-combinaties (bijvoorbeeld zeldzame huishoudsamenstellingen) foutmarges ontstaan. Octavius gaat hier pragmatisch mee om: indien nodig worden rijen met lage plausibiliteit uitgesloten of samengevoegd. Dit leidt soms tot geringe margeschendingen in specifieke subgroepen, maar op het geheel blijven deze binnen acceptabele grenzen.

Op de vraag of Octavius ooit extern is gevalideerd met CBS-microdata, antwoordde Goudappel ontkennend. Wel zijn in enkele projecten intern validaties uitgevoerd, bijvoorbeeld bij backcast-trajecten voor regionale modellen of bij kalibratie van verplaatsingsmodellen op basis van synthetische populaties.

In alle gevallen gold dat interne validatie via margecontrole dominant was, en dat toetsing aan externe observaties beperkt bleef tot verkennende analyses of expertbeoordeling.

Gebruik en toepasbaarheid
Octavius en de onderliggende Population Synthesizer zijn ontwikkeld als generieke, herbruikbare modules voor populatiesynthese binnen verkeersmodellen, maar zijn breder toepasbaar. In het gesprek werd stilgestaan bij de implementatie, de huidige toepassingen en de potentiële inzet buiten het domein van verkeer en vervoer.

De tools zijn ontwikkeld als zelfstandige Java-modules, los van specifieke modellen of platformen. Dit betekent dat de synthesemodule onafhankelijk kan draaien en via een interface gekoppeld wordt aan bijvoorbeeld LMS/NRM, Omnitrans, andere Bentley-platforms, of andere tools voor beleidsanalyse of simulatie.

Deze modulaire architectuur is bewust gekozen om hergebruik te bevorderen en om migratie naar andere modelomgevingen eenvoudiger te maken.

Octavius wordt op dit moment gebruikt in projecten. De tool is geschikt voor zowel trip-based als tour-based modellen, mits de benodigde randdata aanwezig zijn. Voor activity-based modellen is aanvullende logica nodig, met name op het vlak van gedragstoewijzing (allocation), maar de synthese zelf is inzetbaar als basis.
De synthese is doorgaans het eerste element in een modelketen: de populatie wordt gegenereerd, en vervolgens worden kenmerken als voertuigbezit, activiteiten-schema's of verplaatsingen toegekend via andere modules.

Goudappel geeft aan dat in de praktijk veel aandacht uitgaat naar de opbouw van de populatie en dat de kwaliteit van deze stap sterk bepalend is voor het gedrag van het model als geheel.

Hoewel Octavius is ontworpen voor verkeerstoepassingen, is het niet domeinspecifiek. Zolang de methode populaties moet genereren op basis van categorische variabelen en randtotalen, is het systeem ook inzetbaar in andere domeinen, zoals ruimtelijke ordening (bv. modellering van woningvraag), gezondheidszorg (bv. simulatie van zorgbehoefte), arbeidsmarktmodellen (bv. toewijzing van beroepen of participatiegroepen), of onderwijsplanning (bv. leerlingstromen op basis van huishoudsamenstelling en regio).

Deze bredere toepasbaarheid vereist vaak wel aanpassing van variabelen en scenario-invoer, maar niet van de syntheselogica zelf.

Octavius ondersteunt de koppeling met externe variabelen en modellen. Zo kan bijvoorbeeld voertuigbezit als externe input worden meegegeven aan de synthese-module, waarbij deze karakteristiek dient als randvoorwaarde of stuurvariabele. Ook de uitbreiding met modules voor vertrektijd, activiteitengeneratie of tourvorming is technisch mogelijk, en wordt deels al ontwikkeld.

De synthesemodule biedt zo meer dan alleen fitting: het wordt een bouwsteen in een bredere simulatie-architectuur.

Toekomst, openheid en samenwerking
Het gesprek met Goudappel ging ook in op de bredere toekomst van populatiesynthese in Nederland, de wenselijkheid van open source ontwikkeling, en de rol van leveranciers in het organiseren van samenwerking. Daarbij werd ingegaan op de balans tussen innovatie, hergebruik en marktdynamiek.

Goudappel sprak de wens uit om in Nederland toe te werken naar een gemeenschappelijke, open en herbruikbare basis voor populatiesynthese. Niet als één allesomvattend systeem, maar als een modulair platform, waarin verschillende stappen (zoals fitting, allocation, gedragstoewijzing) via gestandaardiseerde interfaces aan elkaar kunnen worden gekoppeld.

Volgens Goudappel is het niet wenselijk dat elke organisatie of regio opnieuw begint met eigen scripts of tools. In plaats daarvan zou populatiesynthese een gemeenschappelijk fundament moeten vormen, net zoals basisregistraties dat zijn voor statistiek of planvorming.

De Population Synthesizer zoals die nu is geïmplementeerd, kan dienen als basis. Maar open source is daarvoor essentieel.

Hoewel de huidige tools (zoals Octavius) niet open source zijn, vindt Luuk dat de richting daar wel naartoe moet. Idealiter zou een generieke open source tool worden ontwikkeld waar heel Nederland gebruik van kan maken, en waarin verschillende methoden en technieken kunnen worden opgenomen. Dit vraagt wel om heldere afspraken over eigenaarschap (wie onderhoudt de code?), governance (wie bepaalt de richting van ontwikkeling?), en financiering (wie betaalt updates, validatie en beheer?).

Goudappel benadrukte dat er al veel kennis en code beschikbaar is, maar dat de fragmentatie in de praktijk leidt tot inefficiënties.

De beweging naar open source vereist ook een andere rol van softwareleveranciers. Partijen als Dat.mobility, Significance en PTV zich meer moeten profileren als platformbeheerders in plaats van leveranciers van black-box tools. Dat betekent een focus op robuuste en onderhoudbare implementaties, ondersteuning bij het configureren en toepassen van de tools, maar minder afhankelijkheid van licenties of exclusieve toegang.

Jan vulde aan dat het marktvolume in Nederland te klein is voor veel concurrerende commerciële pakketten, en dat samenwerking op het vlak van methodiek daarom noodzakelijk is.

Goudappel erkent dat deze overgang gevolgen heeft voor het huidige businessmodel van leveranciers. Toch ziet men voldoende kansen voor nieuwe rollen:

- leveranciers kunnen bijvoorbeeld onderhoud, hosting of kwaliteitsborging van open source tools verzorgen;
- ze kunnen zich toeleggen op gebruiksvriendelijke interfaces, scenario-opbouw of modelkalibratie;
- en ze kunnen met overheden afspraken maken over open standaardmodules met publieke financiering.

Uiteindelijk draait het erom dat populatiesynthese geen ‘product’ is, maar een algemene infrastructuur binnen modeltoepassingen. Dat vereist gezamenlijke investering, maar levert op termijn winst in kwaliteit en consistentie.

Lessen en aanbevelingen

De belangrijkste les is volgens Goudappel eenvoudig maar essentieel: *“Het wiel niet opnieuw uitvinden.”* Veel populatiesyntheseprojecten beginnen met het ontwikkelen van een nieuwe tool, terwijl er al tientallen jaren ervaring is met IPF, IPU, optimalisatie en combinaties daarvan. Goudappel adviseert om eerst bestaande methoden, tools en ervaringen in kaart te brengen, en vervolgens te kiezen voor een beproefde aanpak die modulair uitbreidbaar is. Het is beter om klein en betrouwbaar te beginnen, dan groots en speculatief.

Een andere les is het belang van modulariteit. Door onderdelen zoals *fitting*, *allocation*, *discretisatie*, *validatie*, *gedragstoewijzing* en *scenario-invoer* afzonderlijk te modelleren, ontstaat een flexibel systeem dat beter te beheren is, eenvoudiger te testen en te valideren is, en makkelijker in andere modellen of contexten te gebruiken is. Octavius is volgens dit principe opgebouwd: de synthesemodule staat los van de modelomgeving en kan worden hergebruikt in andere platforms.

Voor toepassingen waar stabiliteit, reproduceerbaarheid en transparantie belangrijk zijn (zoals beleidsmodellen), is het raadzaam om te starten met deterministische methoden. Stochastische technieken (zoals sampling of deep learning) zijn interessant, maar pas in tweede instantie en met voldoende validatie.

Voor SIVMO kan een combinatie van de volgende onderdelen een eerste stap zijn:
- deterministische entropiemaximalisatie (zoals in PopulationSim),
- integerisatie (zoals in Octavius),
- en modulaire scenario-invoer.

Goudappel wees erop dat validatie in de praktijk vaak beperkt blijft tot het controleren van marginale verdelingen. Hoewel dit een noodzakelijke eerste stap is, is het geen garantie voor gedragseigenschappen op micro-niveau. Als bijvoorbeeld huishoudens synthetisch samengesteld zijn met plausibele marges, wil dat nog niet zeggen dat hun voertuigbezit of mobiliteit realistisch is.

Een goede praktijk is validatie op marges (invoer) én gedrag (uitvoer), eventueel door vergelijking met microdata of observaties uit andere bronnen. Goudappel erkende dat dit in projecten vaak beperkt blijft door tijd, geld of opdrachtgeverseisen.

Tot slot adviseert Goudappel om de tool niet alleen af te stemmen op het model van vandaag, maar ook op de vragen van morgen:
- *Schaalbaarheid* naar nationale modellen of microsimulaties;
- *Flexibiliteit* in scenario-opbouw (zoals demografische verandering);
- *Toepasbaarheid* buiten verkeer en vervoer.

Dit vraagt niet om méér complexiteit, maar om een doordachte basisarchitectuur.

3.6 Significance

Methodologische reflectie: synthese vs. transitie
Eén van de kernpunten die Signficance naar voren bracht in het interview, is het belang van het onderscheiden van twee fundamenteel verschillende benaderingen binnen populatiegeneratie:

- Populatiesynthese, waarbij een populatie wordt gegenereerd voor een specifiek basisjaar op basis van waargenomen data (zoals CBS microdata of statistieken).
- Transitiemodellen, waarbij de populatie wordt gesimuleerd over opeenvolgende jaren, waarbij demografische en sociale veranderingen zoals geboorte, sterfte, migratie, opleidingsloopbanen en huishoudvorming expliciet worden meegenomen.

Hoewel het rapport beide benaderingen noemt, vond Significance dat het onderscheid sterker gemaakt zou moeten worden, zowel conceptueel als methodologisch.

Voor het basisjaar van een verkeers- en vervoersmodel is een populatiesynthese op basis van waargenomen marges en microdata (bijvoorbeeld via IPF, IPU of entropie-optimalisatie) volgens Significance de aangewezen methode. Het biedt een gecontroleerde manier om een plausibele, consistente populatie op te bouwen die aansluit bij beschikbare statistieken. Significance merkte op dat de nadruk in het conceptrapport tot nu toe terecht ligt op deze toepassingen, vooral in combinatie met trip-based en tour-based modellen.

Voor langetermijnscenario's, zoals verkenningen richting 2040 of 2050, is volgens Significance bij voorkeur een andere aanpak nodig. Een synthetische populatie voor een toekomstjaar zou dan niet opnieuw gesynthetiseerd moeten worden via marges, maar juist voort moeten vloeien uit een dynamisch transitieproces. Hierbij wordt de bevolking "doorgegroeid" op basis van plausibele aannames over levensloop-gebeurtenissen.

Dit type model biedt een aantal voordelen:

- Consistentie over de tijd (geen plotselinge sprongen in samenstellingen);
- Inherente plausibiliteit (individuen verouderen, veranderen van baan, vormen gezinnen, enz.);
- Meer realistische correlaties tussen kenmerken (bijvoorbeeld inkomen en huishoudenstype).

Hij noemde dit een meer "endogene" benadering van populatie-evolutie, die qua structuur sterk verschilt van de klassieke populatiesynthese. Significance stelde voor om dit onderscheid expliciet te verwerken in het rapport. Bijvoorbeeld door twee duidelijke routes te schetsen. Een synthese van een populatie op tijdstip t_o en een transitie van een populatie van t_o naar $t_1 ... t_n$.

In de huidige tekst zit dat onderscheid volgens hem impliciet verweven bij de beschrijving van simulatieve/generatieve modellen, maar dat is niet voor iedere lezer even duidelijk. Explicitering zou het rapport sterker maken, zeker gezien de beleidsvragen waarvoor SIVMO gebruikt zal worden.

Ervaringen met transitiemodellen in Vlaanderen
Significance gaf een terugblik op de toepassing van transitiemodellen in de Vlaamse modelpraktijk, en de afwegingen die daarbij zijn gemaakt tussen complexiteit, onderhoudbaarheid en beleidsrelevantie. Dit model simuleerde de populatie-ontwikkeling over tijd op basis van sterfte, geboorte, migratie (zowel binnenlands als internationaal), huishoudvorming en echtscheiding, leeftijdsafhankelijkheid, en loopbanen.

Het model had als uitgangspunt dat de gehele populatie “doorgroeit” over de tijd, waarbij agenten (personen en huishoudens) hun eigenschappen behouden, aanpassen of vernieuwen via vooraf gedefinieerde kansen en regels. Hoewel dit model methodologisch zeer volledig was, bleek de praktische implementatie zwaar en onderhoudsgevoelig. De afhankelijkheden tussen deelmodules maakten het lastig om het model op onderdelen aan te passen. Het herstarten van simulaties was complex Het model raakte uiteindelijk buiten gebruik omdat de ontwikkelkosten en het onderhoud niet opwogen tegen het gebruiksgemak en de flexibiliteit.

In de daaropvolgende jaren is Vlaanderen overgestapt naar een ‘lichtere’ aanpak, waarbij per scenariojaar (bijvoorbeeld 2025, 2030, 2040) opnieuw een synthetische populatie wordt gegenereerd met behulp van optimalisatietechnieken. Deze aanpak is eenvoudiger in gebruik en beter beheersbaar, maar heeft wel een belangrijk nadeel: het ontbreken van interjaar-consistentie.

Significance merkte op dat dit in toenemende mate als beperkend wordt ervaren, vooral bij toepassingen waar veranderingen over tijd relevant zijn. Het telkens synthetiseren van losse momentopnames leidt tot “snapshots” die geen logische verbinding hebben, bijvoorbeeld, huishoudens verdwijnen of verschijnen willekeurig, leeftijdsstructuren zijn inconsistent, en mobiliteitsgedrag is moeilijk in te passen in langetermijnlogica.

Het belangrijkste voordeel van transitiemodellen is de continuïteit op micro-niveau. Het modelleert personen die verouderen, die verhuizen, die gezinnen vormen en die veranderen van werk of opleiding. De evolutie van de populatie is daardoor meer realistisch en uitlegbaar. Er ontstaat zodoende een natuurlijk verband tussen scenario’s op verschillende momenten in de tijd.

Voor toepassingen waarin life events, ingroeieffecten of gedragsverandering over de tijd een rol spelen, zijn transitiemodellen volgens hem een sterke toevoeging – mits de technische en organisatorische lasten beheersbaar blijven.

Nadelen van transitiemodellen
Hoewel Significance de inhoudelijke voordelen van transitiemodellen onderstreepte, was men ook duidelijk over de praktische beperkingen en de redenen waarom Vlaanderen op dit moment geen actieve transitietool meer gebruikt voor bevolkingsprojecties in verkeer- en vervoersmodellen.

Een van de grootste nadelen betreft de complexiteit en kosten van transitiemodellen:

- De opbouw van verschillende samenwerkende modules (zoals geboorte, sterfte, migratie en huishoudvorming) vraagt een aanzienlijke investering in ontwikkeling.

- Het model is lastig te onderhouden en aan te passen, vooral bij gewijzigde aannames of databronnen.
- Technische complexiteit leidt tot een grote afhankelijkheid van specifieke ontwikkelteams of consultants, wat de duurzaamheid beperkt.

Significance gaf aan dat ondanks de ambitieuze opzet in 2017, het model slechts tweemaal is gebruikt voor scenarioanalyses. De simulaties werden ingezet voor beleidsverkenningen, maar zijn daarna niet structureel herhaald. Dit suggereert dat de operationele waarde voor beleid op dat moment beperkt was, mede vanwege de grote inspanning die nodig is om het model te draaien en te interpreteren.

Binnen het verkeersdomein blijkt de vraag naar transitiemodellen gering, om meerdere redenen. Veel verkeersmodellen zijn tweepunts-modellen, gericht op een start- en eindjaar (bijvoorbeeld 2025–2050), zonder tussenliggende dynamiek. De inzet van complexe micro-evolutiemodellen wordt daardoor niet altijd als noodzakelijk gezien. Verder zijn beleidsmakers doorgaans geïnteresseerd in vergelijkingen tussen scenario's op vaste momenten, niet in het volledige pad daartussen.

Tot slot benoemde Significance de veranderde institutionele inbedding van het transitiemodel. De verantwoordelijkheid ligt inmiddels bij Statistiek Vlaanderen / afdeling Demografie. Dit heeft als gevolg dat sectorale beleidsmakers, zoals in het verkeer- en vervoersdomein, afhankelijk zijn van andere afdelingen voor het verkrijgen of aanpassen van bevolkingsprojecties. De aansluiting tussen sectorale behoeften (zoals vervoer) en het beheer van het transitiemodel is daarmee minder direct.

Relatie met verkeers- en vervoersmodellen
Significance gaf in het gesprek een reflectie op de relevantie van transitiemodellen voor verkeer- en vervoersmodellen. Hoewel het directe gebruik in transportmodellen beperkt is, kunnen transitiemodellen een belangrijke ondersteunende rol spelen, vooral bij het verrijken van de demografische input.

De meeste verkeersmodellen in Vlaanderen en Nederland (zoals LMS, NRM of BasGoed) zijn scenario-gebaseerd. Ze rekenen met vaste projectiejaren zoals 2025, 2035 en 2050. Er is zelden sprake van simulaties per jaar of met continue ontwikkeling van huishoudens of individuen. Daardoor worden populaties voor elk scenariojaar apart gesynthetiseerd, meestal via ophoging van microdata met margedoelen (IPF, optimalisatie).

Toegevoegde waarde van transitiemodellen
Transitiemodellen kunnen waarde toevoegen door tussenliggende veranderingen en interacties explicieter te maken. Significance noemde meerdere mogelijkheden:
- *Demografische rijkheid*: zoals veranderende huishoudstructuren, opleidingsniveaus, migratieachtergronden en participatiegraden.
- *Arbeidsmarkt- en onderwijsscenario's*: transitiemodellen kunnen doortrekken hoe beleidsmaatregelen (bijv. langere studieduur, pensioenhervorming) zich vertalen naar bevolkingskenmerken.
- *Gender- en generatie-effecten*: hogere deelname van vrouwen aan het hoger onderwijs vertraagt de gezinsvorming, wat leidt tot veroudering van de populatie.

Zulke verschuivingen beïnvloeden bijvoorbeeld het autobezit, mobiliteitsbehoefte en tijdstip van verplaatsingen.

Deze endogene consistentie tussen demografie en gedrag is lastig te realiseren met klassieke populatiesynthese, omdat daar populaties worden geconstrueerd op basis van randverdelingen zonder historische doorwerking.

Significance benadrukte dat rechtstreekse integratie van transitiemodellen in verkeers- en vervoersmodellen vaak te zwaar is. Maar een indirecte koppeling is haalbaar. Men denkt aan eerst simulatie van demografische veranderingen met een transitiemodel. Vervolgens selectie of synthese van populaties voor specifieke scenariojaren op basis van die uitkomsten. Daarna doorvertaling naar transportmodellen via activity/tour/trip-generation.

Op die manier kunnen verkeersmodellen profiteren van de rijkere dynamiek van transitiemodellen zonder dat deze volledig geïntegreerd hoeven te worden.

Alternatief: synthese op basis van externe prognoses
In het gesprek werd ook stilgestaan bij de huidige dominante praktijk: het synthetiseren van populaties voor toekomstjaren op basis van externe demografische projecties van instanties zoals het CBS of commerciële bureaus zoals ABF.

Geconstateerd werd dat veel toekomstscenario's in verkeersmodellen worden gevoed vanuit data zoals:
- *Cohortprognoses* van het CBS, zoals bevolkingsgroei per regio of leeftijdsklasse.
- *Huishoudensprognoses* van ABF, die vaak worden gebruikt in woningbouwplannen.
- *Arbeidsmarkt- of onderwijsprognoses* uit sectorale plannen.

Deze bronnen bieden betrouwbare marges op macroniveau (bijv. aantal 65-plussers in 2050), maar zijn vaak beperkt in microkarakteristieken zoals opleiding, migratieachtergrond, arbeidsparticipatie of mobiliteitskenmerken. Verder zijn ze onderling niet afgestemd, waardoor combinaties van marges kunnen leiden tot inconsistente of onrealistische microdata bij synthese.

Een belangrijk nadeel van deze aanpak is dat populaties voor verschillende scenariojaren los van elkaar worden gesynthetiseerd. De populatie van 2025 en die van 2050 kunnen fundamenteel van samenstelling verschillen, zonder enige interjaarconsistentie. Dit maakt het lastig om ontwikkelingen zoals vergrijzing, migratie-invloeden of onderwijsverschuivingen logisch door te trekken. Ook life events (zoals trouwen, verhuizen, loopbaanontwikkeling) worden niet als transities gemodelleerd, maar slechts als verschillen in marges.

Hoewel deze benadering praktisch is en goed aansluit op bestaande planningspraktijken, mist ze de interne logica en consistentie van een transitiemodel. Elke populatie is een 'snapshot' die apart is gegenereerd. Er is geen natuurlijke relatie tussen de microdata van jaar t en jaar t+n. Dit kan tot incoherente of discontinu populatie-invoer leiden in modellen die dynamiek of gedrag willen modelleren.

Significance stelde dat dit een wezenlijke beperking vormt voor toepassingen met hogere eisen aan interne consistentie, zoals agent-based modellen, maar dat het in veel **7.**

Samenvattende les
Aan het eind van het gesprek vatte Significance hun advies aan het SIVMO-project kernachtig samen, gebaseerd op hun ervaring met zowel synthese- als transitiemodellen in Vlaanderen. Voor het opstellen van een realistische en valide populatie voor het basisjaar is populatiesynthese de meest geschikte methode. Zeker wanneer er voldoende microdata beschikbaar zijn, er betrouwbare randtotalen (marges) beschikbaar zijn en de populatie dient als input voor verkeers- of vervoersmodellen. Populatiesynthese biedt in deze context stabiliteit, reproduceerbaarheid en controle over de output.

Als de focus ligt op toekomstige ontwikkelingen, en vooral op dynamiek tussen jaren (bijv. ontwikkeling van huishoudstructuur of opleidingsniveau), consistentie binnen individuen of huishoudens over de tijd, en gedragsmatige toepassingen zoals agent-based modellering, dan kan een transitiemodel een betere oplossing zijn. Het biedt de mogelijkheid om veranderingen logisch en samenhangend te simuleren, in plaats van gesynthetiseerde snapshots per jaar.

Significance benadrukte echter ook het belang van pragmatiek. Transitiemodellen zijn duur, complex en onderhoudsgevoelig. Als het model slechts wordt gebruikt om één of twee toekomstscenario's door te rekenen, is het de investering vaak niet waard. De initiële ontwikkelkosten kunnen niet worden terugverdiend zonder herhaald gebruik.

Tot slot wees Significance op het belang van structureel eigenaarschap. Als een transitiemodel te specialistisch wordt, is beheer door één beleidsdomein (zoals verkeer) meestal niet haalbaar. In Vlaanderen is het beheer daarom verschoven naar de afdeling Statistiek/Demografie van de Vlaamse overheid. Sectoren zoals verkeer zijn dan afhankelijk van externe partijen voor updates, gebruik en aansluiting op hun modellen.

Voor SIVMO betekent dit dat een keuze voor transitiemodellen ook organisatorische implicaties heeft — niet alleen technisch, maar ook qua governance en beheer.

Bijlage 4 Lijst met afkortingen

- **AcBM** – *Activity-Based Model*: model mobiliteitsgedrag van personen op basis van activiteitenpatronen bepaald.
- **AgBM** – *Agent-Based Model*: simulatiemodel waarin het gedrag van individuele agenten (zoals voertuigen, personen of huishoudens) wordt nagebootst.
- **AIPF** – *Adjusted Iterative Proportional Fitting*: variant op IPF met aangepaste correctiestappen om convergeerproblemen te verminderen.
- **CART** - *Classification and Regression Trees*; een boomgebaseerde methode die data opsplitst in steeds homogener groepen om een klasse te voorspellen (classificatie) of een continue waarde te schatten (regressie).
- **COS** – *Combinatorial Optimisation Sampling*: populatiesynthese via heuristische bewerkingen zoals toevoeging/verwijdering van huishoudens.
- **DGP** – *Data Generating Process*: het onderliggende proces dat data genereert; van belang voor validatie van synthetische populaties.
- **GAN** – *Generative Adversarial Network*: deep learning-techniek waarbij twee neurale netwerken synthetische data genereren en beoordelen.
- **GTFS** – *General Transit Feed Specification*: open standaard voor openbaarvervoerdata, o.a. gebruikt in MATSim.
- **GUI** – *Graphical User Interface*: visuele gebruikersomgeving van software.
- **INNLS –** *Iterative non-negative least squares:* een iteratieve rekenmethode die per microdata-record een niet-negatief gewicht bepaalt, zodat de opgetelde gewichten per categorie zo dicht mogelijk aansluiten op de opgelegde marges (randtotalen).
- **IPF** – *Iterative Proportional Fitting*: methode om een steekproefverdeling stapsgewijs aan te passen aan marginale totalen.
- **IPU** – *Iterative Proportional Updating*: uitbreiding van IPF waarbij gelijktijdige aanpassing op huishouden- én persoonsniveau mogelijk is.
- **KL-divergentie** – *Kullback–Leibler-divergentie*: statistische maat voor verschil tussen twee kansverdelingen; gebruikt in optimalisatie.
- **LMS** – *Landelijk Model Systeem*: Nederlands macroscopisch verkeersmodel beheerd door Rijkswaterstaat.
- **MATSim** – *Multi-Agent Transport Simulation*: open source agent-based simulatiemodel met eigen populatiesynthesemodule.
- **MCMC** – *Markov Chain Monte Carlo*: Rekenmethode waarmee steekproeven worden genomen uit kansverdelingen via afhankelijke trekkingen.
- **MON** – *Mobiliteitsonderzoek Nederland*: Enquêtes over mobiliteit in Nederland.
- **MPN –** *Mobiliteitspanel Nederland*: Panelonderzoek naar mobiliteit.
- **MWCOG** – *Metropolitan Washington Council of Governments*: Amerikaanse planningsinstantie, gebruiker van PopulationSim.
- **NNLS –** *Non-negative least squares:* een optimalisatiemethode die niet-negatieve gewichten schat zodat een combinatie van records zo goed mogelijk aansluit op de randtotalen.
- **NRM** – *Nederlands Regionaal Model*: regionaal verkeersmodel als aanvulling op het LMS.
- **NTEM** – *National Trip End Model*: geaggregeerd trip-end model uit het VK.
- **NTS** – *National Travel Survey*. Mobiliteitsonderzoek in het VK.

- **ODiN** – *Onderweg in Nederland*: nationale mobiliteitsenquête van het CBS.
- **OVG** – Onderzoek Verplaatsingsgedrag: Enquêtes over mobiliteit in Nederland.
- **PoPuS** – *Population Synthesis*: algemene afkorting gebruikt in Engelstalige literatuur.
- **RA** – *Remote Access:* Omgeving van het CBS, waarin geautoriseerde gebruikers onder strikte voorwaarden met microdata kunnen werken zonder dat de ruwe microdata de beveiligde omgeving verlaat.
- **RMSE** – *Root Mean Square Error*: maat voor gemiddelde afwijking; gebruikt bij validatie.
- **SNET –** *Statistical Noise Elimination Technique;* een deterministische techniek om continue gewichten om te zetten naar gehele aantallen (integers), met beperkte discretisatiefout en zonder stochastische sampling.
- **SIVMO** – *Samenwerkingsverband en Innovatie Verkeersmodellen door Overheden;* opdrachtgever voor deze studie.
- **TAZ** – *Traffic Analysis Zone*: ruimtelijke eenheid voor verkeer- en vervoermodellen, o.a. in de VS.
- **VAE** – *Variational Autoencoder*: deep learning-techniek om complexe verdelingen te modelleren op basis van latente variabelen.

Bijlage 5 Begrippenlijst

- **Agent**: Een individu of huishouden dat zelfstandig beslissingen neemt in een simulatiemodel (bijv. over reizen of autobezit).
- **Attributen**: Kenmerken van personen of huishoudens, zoals leeftijd, inkomen of autobezit, die worden gebruikt om populaties te specificeren.
- **Backcast**: Terugrekening van modeluitkomsten in het verleden, gebruikt om te controleren of een model realistische uitkomsten geeft.
- **Control totals / marges**: Externe, vaak geaggregeerde gegevens (zoals het aantal personen per leeftijdsklasse) waaraan een synthetische populatie moet voldoen.
- **Datafusie**: Het combineren van verschillende databronnen (zoals enquêtes en registerdata) tot één consistente dataset.
- **Deterministisch model**: Model waarbij eenzelfde invoer altijd exact dezelfde uitvoer oplevert, zonder toeval.
- **Entropiemaximalisatie**: Optimalisatieprincipe waarbij een verdeling wordt gekozen die zo min mogelijk aanname introduceert, gegeven bekende marges.
- **Integerisatie**: Stap waarbij continue of gewogen resultaten (bijv. 1,3 huishoudens) worden omgezet naar gehele aantallen (bijv. 1 of 2 huishoudens).
- **Kalibratie**: Proces waarbij een model wordt afgestemd op bekende gegevens, zodat de uitkomsten overeenkomen met waarnemingen.
- **Kullback–Leibler-divergentie**: Statistische maat voor het verschil tussen twee verdelingen; kleiner betekent betere overeenkomst.
- **Lege cellen (‘zero-cell problem’)**: Combinaties van kenmerken die niet voorkomen in de steekproef maar wel verwacht worden in de populatie.
- **Marginale distributie**: Verdeling van één variabele (zoals leeftijd) los van andere kenmerken, vaak beschikbaar uit censusdata.
- **Microsimulatie**: Modellering waarbij gedrag en kenmerken van individuele eenheden (personen/huishoudens) worden gevolgd door de tijd.
- **Optimalisatie**: Rekentechniek waarbij een oplossing wordt gezocht die het best voldoet aan randvoorwaarden, zoals marges of minimumnormen.
- **Populatiesynthese**: Het genereren van een kunstmatige maar realistische populatie van personen of huishoudens op basis van steekproefdata en marges.
- **Reproduceerbaarheid**: Eigenschap van een methode waarbij dezelfde invoer altijd tot exact dezelfde uitkomsten leidt.
- **Sampling**: Steekproeftrekking waarbij op basis van kansen records uit een databestand worden gekozen, vaak met toeval.
- **Seed dataset**: De oorspronkelijke (steekproef)dataset die als basis dient voor populatiesynthese.
- **Simulatief model**: Model dat gedrag of ontwikkelingen nabootst via iteratieve of stochastische processen, vaak over de tijd.
- **Synthpop / synthetische populatie**: Een kunstmatig samengestelde groep individuen of huishoudens die qua structuur lijkt op de echte populatie, maar geen echte personen bevat.
- **Validatie**: Proces waarbij gecontroleerd wordt of de uitkomsten van een populatiesynthese overeenkomen met externe gegevens of verwachte patronen.
- **Weighting / herweging**: Techniek om records in een steekproef zwaarder of lichter te laten meetellen, zodat de verdeling beter past bij bekende marges.