Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Eidgenössisches Departement für
Umwelt, Verkehr, Energie und Kommunikation UVEK

Bundesamt für Raumentwicklung ARE
Sektion Grundlagen

EBP

Workshop-Bericht: Synthetische Populationen für die Politikberatung in der Schweiz

Bern, 8. Dezember 2017

Bundesamt für Raumentwicklung ARE, Sektion Grundlagen
Nicole Mathys
Andreas Justen

EBP
Peter de Haan
Adrian Stetter

30. Januar 2018

Vorwort

Das Bundesamt für Raumentwicklung ARE hat die Aufgabe Grundlagen und Strategien in den Bereichen Raumentwicklung und Gesamtverkehr zu entwickeln. Wichtige Arbeitsinstrumente sind dabei Modelle welche die Bevölkerungs- und Arbeitsplatzentwicklung (Flächennutzungsmodelle) sowie die verkehrlichen Entwicklungen (Verkehrsmodelle für den Personen- und Güterverkehr) abbilden. Das ARE hat im Eidgenössischen Departement für Umwelt, Verkehr, Energie und Kommunikation UVEK die Aufgabe diese Modelle zu erstellen, auf neue Inputdaten zu aktualisieren und methodisch weiter zu entwickeln. In diesem Zusammenhang hat das ARE erste Erfahrungen in der Anwendung von Synthetischen Populationen (SynPop) gemacht.

Unter einer SynPop wird ein aus verschiedenen Datenquellen zusammengestellter Datensatz zu Personen und Haushalten, der vielfache demografische und soziökonomische Attribute enthält (z.B. Alter, Geschlecht, Bildungsstand, Einkommen, Verfügbarkeit ÖV-Abonnemente und Personenwagen) verstanden.

Da die Verwendung von SynPop eine methodische Neuigkeit darstellt hat das ARE zu einem Workshop eingeladen. Verschiedene Akteure, sowohl Ersteller als auch aktuelle und potenziellen Nutzer von SynPop, aus der Forschung, aus Beratungsbüros und der Verwaltung der Kantone und des Bundes haben Informationen ausgetauscht und debattiert. Ziel des Zusammentreffens war es, eine aktuelle und umfassende Übersicht über bestehende Bedürfnisse und Anwendungen zu erhalten, im Gespräch gegenseitig voneinander zu lernen und Empfehlungen für die Weiterentwicklung und Anwendung von SynPop zu formulieren. Der vorliegende Tagungsbericht führt in das Thema der SynPop ein, hält die Inhalte der Präsentationen fest, und gibt Empfehlungen für die Erstellung und die Anwendung von SynPop ab.

Der Workshop stoss auf reges Interesse, die Teilnehmerliste finden Sie in Kapitel 5. Ein kurzer Fragebogen zur Evaluation am Ende des gemeinsamen Tages zeigte auch auf, dass ein Bedürfnis zum weiteren Austausch besteht. Wir danken für die regen und offenen Diskussionen am Workshop und werden auch in Zukunft aktiv zum Austausch in der SynPop-Community beitragen. Ein zweiter Workshop ist voraussichtlich für Ende 2019 geplant.

Andreas Justen

Nicole Mathys

Sektion Grundlagen

Bundesamt für Raumentwicklung ARE

Inhaltsverzeichnis

1. Einleitung

1.1 Ausgangslage

Die Tradition der Volkszählung reicht in der Schweiz weit zurück. Seit 1850 lieferte die klassische Volkszählung alle zehn Jahre wichtige Informationen über die Bevölkerung, Haushalte, Gebäude und Wohnungen. Im Jahr 2000 fand die letzte Volkszählung dieser Art statt. Seit 2010 setzt das BFS die *Neue Volkszählung* um. Um die Bevölkerung zu entlasten, werden viele Informationen aus den vorhandenen Einwohner-, Betriebs-, Gebäude- und Wohnungsregistern der Gemeinden und Kantone entnommen. Diese Daten werden mit Stichprobenerhebungen ergänzt. Dabei wird nur ein kleiner Teil der Bevölkerung schriftlich oder telefonisch befragt. Die Schweiz verfügt damit über ein modernes statistisches System, welches die Strukturen und die Entwicklung der Bevölkerung, Haushalte, Betriebe, Gebäude und Wohnungen kontinuierlich beobachtet. Die in die *Neue Volkszählung* einfliessenden Informationen basieren dabei auf folgenden vier Kernelementen:

1. Die **Registererhebung** liefert grundlegende Informationen zum Bestand und zur Struktur der Bevölkerung, Haushalte und Betriebe sowie der Gebäude und Wohnungen. Es handelt sich um *Vollerhebungen* aller Personen und Haushalte bzw. Betriebe (STATPOP[1] bzw. STATENT) sowie der Gebäude inkl. ihrer genauen geographischen Lage (Gebäude- und Wohnungsregister, GWR), welche ständig nachgeführt werden. Unter Berücksichtigung des Datenschutzes werden aus den Registererhebungen durch das BFS anonymisierte statistische Standardprodukte erstellt.

2. Bei der **Strukturerhebung** wird jährlich eine *Stichprobe* von ca. 2.5% (200'000 Personen) schriftlich oder online befragt. Die Erhebung ergänzt die Informationen der Register um Statistiken zu den Themen Bevölkerung, Haushalte, Familie, Wohnen, Arbeit, Mobilität, Bildung, Sprache und Religion.

3. Die **thematischen Erhebungen** werden bei *Stichproben* von 0.1% bis 0.7% der Bevölkerung (10’000- 40'000 Personen) durchgeführt; in jedem Jahr findet eine Erhebung zu einem spezifischen Thema statt, was Periodizitäten von drei (z.B. für die Haushaltbudgeterhebung, HABE) bis fünf Jahre (für den Mikrozensus Mobilität und Verkehr, MZMV) ergibt. Mit den Statistiken dieser Erhebungen können die Informationen aus der Strukturerhebung wesentlich vertieft werden.

4. **Omnibus-Erhebungen** sind telefonische Befragungen einer *Stichprobe* von 3’000 Personen (mindestens einmal jährlich) für die rasche Beantwortung von wechselnden politischen oder wissenschaftlichen Fragestellungen.

Seit 2010 setzt das BFS die *Neue Volkszählung* um, seit ca. 2015 liegen alle Erhebungen in guter, belastbarer Qualität vor. Damit bietet sich neu die Möglichkeit, für wichtige Zusammenhänge die Erhebungen der *Neuen Volkszählung* zu kombinieren zu einem *Gesamtdatensatz*: Wichtige Attribute aus den Stichprobenerhebungen müssen dazu an die Datensätze der Vollerhebungen «angespielt» werden – es entsteht eine *synthetische Population*.

[1] Für weitere Informationen zu diesen Datensätzen siehe Kapitel 1.4

1.2 Was ist eine synthetische Population?

Eine synthetische Population (SynPop) beschreibt einen Datensatz von Bevölkerung und Haushalten, der vielfache demografische und soziökonomische Attribute der Personen und Haushalte vorhält (z.B. Alter, Geschlecht, Bildungsstand, Einkommen, Verfügbarkeit von Mobilitätswerkzeugen wie ÖV-Abonnemente und Personenwagen sowie eine Differenzierung nach Haushaltstypen). Die Besonderheit besteht darin, dass je Person (und Haushalt) die jeweiligen Attribute vollständig vorliegen und der Datensatz damit die Bevölkerung derart detailliert beschreibt, wie dies letztmalig durch die Volkszählung 2000 annähernd möglich war. Aufgrund der sehr guten Datenlage in der Schweiz (siehe Kapitel 1.1 und 1.4) kann die Erstellung einer SynPop auf realen Daten aufsetzen; entsprechend ist die SynPop über die Anreicherung weiterer Datensätze eher als *teilsynthetisch* zu beschreiben. Im Ausland, wo entsprechende Grundlagen nicht verfügbar oder zugänglich sind, sind vollsynthetische Verfahren einzusetzen. Zur Vereinfachung wird im Folgenden von SynPop gesprochen, gemeint sind dabei sowohl teil- als auch vollsynthetische Verfahren.

Die Anwendungsbereiche für eine SynPop sind vielfältig: Je nachdem, welche Attribute die SynPop vorhält, eignet sie sich zur differenzierten Betrachtung von z.B. Fragestellungen im Raumentwicklungs-, Mobilitäts- und Energiebereich. Die Daten sind grundsätzlich auch interessant für Fragen der Marktforschung (z.B. Unternehmensstandortwahl). Eine zusätzliche Inwertsetzung erfährt eine SynPop dann, wenn es möglich ist, die Attribute für einen zukünftigen Zeitpunkt fortzuschreiben.

Da ein solcher Datensatz wie die klassische Volkszählung, nicht (mehr) verfügbar ist, muss er unter Anwendung statistischer Verfahren aus verschiedenen Datenquellen erstellt werden. Damit einher geht eine Einschränkung: Die Analysen zu Zusammenhängen wie auch die Fortschreibung sind hinsichtlich ihrer Validität davon abhängig, wie gut, d.h. realitätsnah es gelingt die Verschränkung der Attribute aus unterschiedlichen Datenquellen zu realisieren.

1.3 Ziele des Workshops vom 8. Dezember 2017

Die Forschung an Methoden zur Erstellung einer SynPop sowie deren Nutzung in Praxisprojekten erfolgt in der Schweiz derzeit an verschiedenen Stellen. Von unterschiedlichen Zielanwendungen und entsprechend divergierenden Zielsetzungen ausgehend, werden die synthetischen Populationen mit verschiedenen Methoden erstellt. Dabei kommen aber meist die gleichen Grunddatensätze zum Einsatz.

Die Heterogenität in Erstellung und Anwendung ist aus methodischer Sicht zu begrüssen, da aktuell unterschiedliche Arbeitsweisen und Ansätze in Konkurrenz zueinanderstehen. Gleichzeitig entstehen Datensätze, die sich aufgrund der Methodik als auch der jeweils in den Fokus gerückten Zielsetzung voneinander unterscheiden, dabei aber gleichermassen als „SynPop“ bezeichnet werden. Für die Nutzer von SynPop-Datensätzen – dazu gehören auch die Bundes- und Kantonsverwaltungen – ist es eine Herausforderung zu bewerten, welche SynPop-Methoden und -Datensätze für eine gegebene Fragestellung jeweils die bestgeeigneten sind.

Der Workshop vom 8. Dezember schafft erstmals in der Schweiz eine Übersicht über die verschiedenen Ansätze, Methoden und verwendeten Grunddatensätze.

Der Workshop verfolgt die folgenden Ziele:

- **Gemeinsames Verständnis und Übersicht:**
  - Die Möglichkeiten und Herausforderungen von SynPop-Anwendungen bekannt machen, insbesondere für Akteure der Bundes- und Kantonsverwaltungen;
  - Gesamtübersicht hinsichtlich der aktuell eingesetzten Daten und den Methoden zu deren Zusammenführung zu einer SynPop;
  - Zukünftige Nutzer sollen die möglichen Ansätze inkl. ihrer Potenziale und Nachteile bewerten können.
- **Best-Practice: Daten, Methoden und Qualitätssicherung**
  - Entwickler von SynPop sollen sich austauschen und wo sinnvoll einen Abgleich der eingesetzten Methoden und Daten einleiten können;
  - Empfehlungen für die Weiterentwicklung der Erstellung und der Anwendung von SynPop sind zu formulieren, und wo sinnvoll sollen Chancen zur Harmonisierung von Vorgehensweisen identifiziert werden;
  - Übersicht über mögliche Ansätze und Integration der Qualitätssicherung in Projektabläufen.
- **Anwendungen:**
  - Neue Anwendungsbereiche, wo bis anhin nur einzelne Grunddatensätze, oder regional definierte Teilstichproben derselben, verwendet werden, sollen identifiziert werden.

1.4 Die Grunddatensätze des BFS

Die folgenden Datensätze des Bundesamts für Statistik bilden die wichtigste Grundlage zur Modellierung der schweizerischen synthetischen Populationen und werden in diesem Bericht als Grunddatensätze bezeichnet.

Diese Register- und Erhebungsdaten sind Eigentum des BFS und nicht des jeweiligen Erstellers oder Anwenders einer synthetischen Population. Für die Datensätze STATPOP und STATENT existieren Lizenzmodelle, welche genereller Natur sind und die Datenverwendung nicht auf konkrete Projekte oder Anwendungen eingrenzen. Diese beiden Erhebungen sind für nichtkommerzielle Anwendungen frei verfügbar, während für kommerzielle Zwecke vorgängig eine Bewilligung beim BFS eingeholt werden muss. Für die übrigen Datensätze benötigt jede Entwicklung und Anwendung einer synthetischen Population fallweise entsprechende Datenlizenzen. Diese sind für Bundesämter und kantonale Behörden in der Regel kostenfrei. Das BFS als Eigentümerin der Grunddatensätze stellt sicher, dass sie nur für die gesetzlich festgelegten statistischen Aufgaben verwendet werden. Bei der Publikation von Daten stellt das BFS sicher, dass keine Einzelpersonen oder Einzelhaushalte identifiziert werden können. Für jeden Datensatz ist mit dem BFS ein Datenvertrag abzuschliessen. Dieser schreibt jeweils auch die Einhaltung der Datenverknüpfungsverordnung (SR 431.012.13) des Eidgenössischen Departements des Innern (EDI) vor.

Statistik der Bevölkerung und der Haushalte – STATPOP	
Stichprobe und Periodizität	Vollerhebung; jährlich
Kurzbeschrieb	Die Statistik der Bevölkerung und der Haushalte ist Teil des eidgenössischen Volkszählungssystems. Sie liefert Informationen zum Bestand und zur Struktur der Wohnbevölkerung am Jahresende sowie zu den Bevölkerungsbewegungen während des Kalenderjahres. Zusammen mit der Strukturerhebung bildet sie zudem die Grundlage für die Haushaltsstatistik.
Link zur BFS-Unterseite	STATPOP

Statistik der Unternehmensstruktur – STATENT	
Stichprobe und Periodizität	Vollerhebung; jährlich
Kurzbeschrieb	Die STATENT liefert zentrale Informationen zur Struktur der Schweizer Wirtschaft (z. B. Anzahl Unternehmen, Anzahl Arbeitsstätten, Anzahl Beschäftigte usw.) und gibt damit einen Überblick über die Wirtschaftslandschaft der Schweiz.
Link zur BFS-Unterseite	STATENT

Eidgenössisches Gebäude- und Wohnungsregister – GWR	
Stichprobe und Periodizität	Vollerhebung; vierteljährlich
Kurzbeschrieb	Das eidgenössische Gebäude- und Wohnungsregister (GWR) ist im Anschluss an die Volkszählung 2000 auf der Grundlage der damaligen Gebäude- und Wohnungserhebung aufgebaut worden und umfasst alle Gebäude mit Wohnnutzung und deren Wohnungen in der Schweiz. Geführt werden neben schweizweit eindeutigen Gebäude- und Wohnungsidentifikatoren (EGID bzw. EWID) die wichtigsten Grunddaten wie Adresse, Standortkoordinaten, Baujahr, Anzahl Geschosse, Heizungsart für die Gebäude sowie Anzahl Zimmer und Wohnungsfläche für die Wohnungen.
Link zur BFS-Unterseite	GWR

Strukturerhebung – SE	
Stichprobe und Periodizität	Mind. 200'000 Personen; jährlich
Kurzbeschrieb	Die Strukturerhebung (SE) ist ein Element der Volkszählung und ergänzt die Informationen aus den Registern mit zusätzlichen Statistiken zur Bevölkerungsstruktur. Dabei wird ein Teil der Bevölkerung schriftlich befragt. Erste Resultate stehen ein Jahr nach dem Stichtag zur Verfügung
Link zur BFS-Unterseite	SE

Mikrozensus Mobilität und Verkehr – MZMV (BFS/ARE)	
Stichprobe und Periodizität	60'000 Personen; CATI; alle 5 Jahre (nächste Durchführung 2020)
Kurzbeschrieb	Der MZMV erfasst die Mobilität der Schweizer Bevölkerung. Erfasst werden alle Wege und Etappen an einem Stichtag (typischerweise der Vortag) sowie längere Reisen in den letzten drei Monaten. Erhoben werden auch die Motorfahrzeuge des Haushalts sowie die ÖV-Abonnemente der Zielperson. Zu jeder Etappe werden Verkehrsmittel, geocodierter Anfangs- und Endpunkt sowie Verkehrszweck erhoben.
Link zur BFS-Unterseite	MZMV

Haushaltsbudgeterhebung – HABE	
Stichprobe und Periodizität	250 Haushalte pro Monat; kontinuierlich
Kurzbeschrieb	Die Haushaltsbudgeterhebung (HABE) hat zum Ziel, die Haushaltsbudgets der Wohnbevölkerung in der Schweiz detailliert zu erfassen. Die teilnehmenden Haushalte notieren während eines Monats alle anfallenden Ausgaben und Einkommen in die Erhebungsdokumente. Sie werden dabei von erfahrenen Spezialisten telefonisch betreut.
Link zur BFS-Unterseite	HABE

2. Entwicklung von synthetischen Populationen

Zu jedem Referat des Workshops bietet dieses Kapitel Kurztexte mit den wichtigsten Inhalten. Die vollständigen Foliensätze finden sich im Anhang.

2.1 Synthetische Populationen aus FaLC-sim für das Nationale Personenverkehrsmodell NPVM (Bodenmann, Strittmatter AG)

Synthetische Population «FalC-sim» für das Nationale Personenverkehrsmodell NPVM	
Ersteller	regioConcept AG, Herisau www.falc-sim.org
Partner	ETH Zürich (Institut für Verkehrsplanung und Transportsysteme, IVT; Institut für Raum- und Landschaftsentwicklung, IRL/PLUS), Imperial College London, Bundesamt für Raumentwicklung, SBB, Strittmatter Partner AG, ESMO Žilina a.s. (Slovakei), datatools GmbH, Fahrländer Partner AG, u.a.m.
Software (+open source: Ja/ Nein)	FaLC-core (Ja), FaLC-synpop (Nein), diverse Skripte in FaLC-skript (Ja)
Themengebiete	Szenarien Flächennutzung und Verkehrsentwicklung (Verteilung Wohnbevölkerung und Arbeitsplätze, Pendlerbewegungen; Auswirkungen von Änderungen der Verkehrsinfrastruktur, Bauzonengrösse, Steuern u.a.m.)
Zeitliche & Geographische Auflösung	Startpopulationen für 2000, 2010, 2014, 2016 Anschliessend jährliche Fortschreibung (veränderbar) NPVM-Zonen in der Schweiz
Nächste Schritte	Erweiterung und Einbezug zusätzlicher Informationen (u.a. Einbezug Kaderlöhne und Teilzeitarbeit, verbesserte Modellierung Landpreise) Abgleich der Synthetischen Population 2030 für die Schweiz mit den Prognosen des Bundesamtes für Statistik (Kalibration auf Stufe Kantone) Synthetische Population 2017
Prognosefähigkeit	Ja, mit Evolutionsmodell, das in Jahresschritten arbeitet
Grösse der grundlegenden Datenmenge	Input-Datenbank: 30 GB Output (Tabellen Personen, Haushalte, Unternehmen): - ganze Schweiz 2.2 GB / 8.4 Mio. Personen - Kantone SG/AR/AI 130 MB / 570'000 Personen

SynPop wozu?

Planer, Behörden, Bauherren, Investoren sowie Einwohner und Firmen beschäftigen sich mit verschiedenen Fragestellungen im Zusammenhang mit Raum und Verkehr. Diese Fragen können einzelne Parzellen oder Gebiete wie Städte, Gemeinden, Kantone oder auch ganze Länder betreffen. Das Ziel einer zeitgemässen Raum- und Verkehrsplanung ist eine integrierte und vorausschauende Entwicklung des Raumes. Bei Strittmatter Partner werden die Informationen aus FaLC und dessen synthetischen Populationen in verschiedensten Projekten genutzt. Insbesondere bei Ortsplanungsrevisionen lohnt sich die Datenaufbereitung und Auswertung, gibt das Modell doch Auskunft über verschiedene planerisch relevante Themen wie Bevölkerungsentwicklung, Schülerzahlen, Alterung, Nachfrage nach Wohn- und Arbeitsflächen u.v.m. Wesentlich dabei ist natürlich gerade auch die Entwicklung in die Zukunft.

FaLC: Simulation von Wohnen – Arbeiten – Verkehr

Ziel von **FaLC «Facility Location Simulation Tool»** ist die Unterstützung von räumlichen Planungsprozessen unter Berücksichtigung der Wechselwirkungen zwischen Bautätigkeit, Migrationen, Pendlerbeziehungen und Verkehrsentwicklung. FaLC ermöglicht für frei wählbare Szenarien die Simulation der (zukünftigen) Entwicklung der Wohnbevölkerung, Haushalte, Arbeitsplätze und Unternehmen auf verschiedenen räumlichen Aggregationsstufen. Dies beinhaltet indes zahlreiche weitere Informationen: z.B. das Alter und Einkommen der Personen, Pendlerbeziehungen sowie die räumliche Verteilung der Unternehmen verschiedener Branchen. Mögliche Fragen, die mit dem Vergleich von verschiedenen Szenarien beantwortet werden können, sind beispielsweise:

- Was sind die räumlichen Effekte einer Veränderung des Verkehrsinfrastrukturangebots auf die Verteilung der Wohnbevölkerung und die Beschäftigten – beispielsweise eines neuen Autobahnanschlusses oder schnelleren Bahnverbindungen?
- Was sind die räumlichen Effekte einer Reduktion der Steuern – beispielsweise, wenn Gemeinde X in der Agglomeration Y die Einkommenssteuern deutlich senkt?
- Was sind die räumlichen Effekte von Änderungen der kommunalen Nutzungsplanung– beispielsweise eine allgemeine oder teilweise Ermöglichung dichterer Bauweisen?

In FaLC werden die Entscheide von einzelnen Personen, Haushalten und Unternehmen abgebildet. Dieser Ansatz ermöglicht ausserordentlich viele Indikatoren, um Effekte und Nebeneffekte sichtbar zu machen. Einige Beispiele sind die Entwicklungen

- der Altersverteilung der Bevölkerung;
- der Pendlerdistanzen;
- der Bautätigkeit und
- des Bodenmarktes.

Zudem berücksichtigt FaLC in den Szenarien exogene Einflüsse wie die Entwicklung der globalen Wirtschaftsentwicklung sowie gesellschaftliche Tendenzen wie die Veränderung der Fertilitätsraten oder

Arbeitszeitveränderungen. Die agentenbasierte Mikrosimulation – d.h. die Modellierung einzelner Personen, Haushalte und Unternehmen – hat den grossen Vorteil, dass die Eigenschaften der modellierten Personen abhängig von den Fragestellungen relativ einfach erweitert werden können.

Synthetische Population in FaLC

Zusätzlich zu den typischen Attributen der Agenten wie Alter, Geschlecht, Haushaltszugehörigkeit, Wohn- und Arbeitsort werden verschiedene zusätzliche Informationen modelliert, wie: Hauptsprache, Bildungsstand, Einkommen und Mobilitätswerkzeuge. Die aktuellen Populationen für das NPVM enthalten die adressscharfen Koordinaten der Haushalte und Unternehmen (metergenau). Je nach Verwendungszweck (und entsprechenden Datenrestriktionen) können aber auch aggregierte Hektar-Rasterdaten verwendet werden.

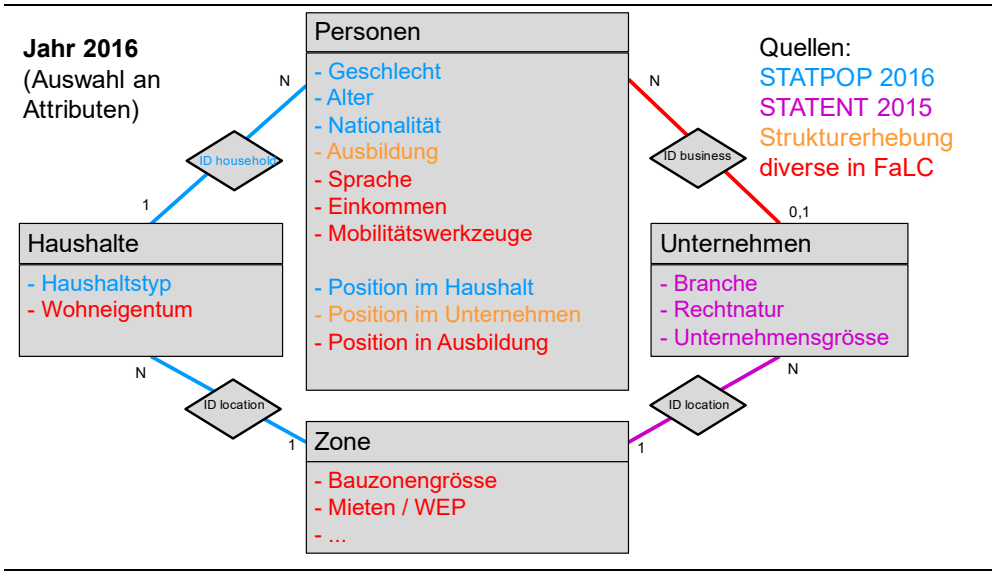


Abbildung 1: Input-Daten zu synthetischen Population in FaLC (WEP=Wohneigentumspreise)

Die meisten imputierten Informationen basieren auf Regressionsmodellen, die die Wahrscheinlichkeiten für bestimmte Ausprägungen beschreiben und werden mit Monte-Carlo-Simulationen modelliert. Die hierfür typischen multinominalen (MNL)-Regressionsmodelle erreichen indes nur sehr tiefe Bestimmtheitsmasse und sind deshalb auf der Stufe der einzelnen Entitäten ungenau. Je nach Ebene für die Kalibration werden die Randsummen auf den Aggregationsstufen unterschiedlich präzise abgebildet. Am präzisesten werden zurzeit die Mobilitätswerkzeuge wie Besitz eines Personenwagens, eines GA- oder Halbtax-Abonnements abgebildet – da diese Informationen direkt auf die Randsummen der einzelnen Zonen des Verkehrsmodells des ARE (NPVM-Zonen) kalibriert werden können. Grundsätzlich weisen die meisten Ausprägungen aber spätestens auf Stufe Planungsregion eine Fehlererwartung von unter +/– 1% aus.

2.2 Wie schweizerisch ist die synthetische Schweiz von EBP? (Müller, EBP)

Synthetische Schweiz von EBP	
Ersteller	EBP
Partner	Kanton Zürich, Amt für Verkehr / SBB
Software	Programmiersprache R (open source)
Themengebiete	Haushalte, Wohnsituation, Mobilitätsverhalten und -Werkzeuge, Konsumverhalten
Zeitliche & Geographische Auflösung	Zeitlich: Ausschlaggebend ist das Datum der verwendeten zugrundeliegenden GWS-Daten; es werden die neuesten verfügbaren Datensätze (HABE, MZVM) angehängt Geographisch: rechnerisch Hektar-Auflösung, Auswertung teilweise nur auf Gemeindeebene sinnhaft
Nächste Schritte	> kontinuierliche Verbesserung der statistischen Verknüpfungsmodelle > Implementierung Zukunftsszenarien
Prognosefähigkeit	Nein, jedoch Implementierung geplant
Grösse der grundlegenden Datenmenge	Teil-synthetische (Echtdatensätze ergänzt um imputierte Stichproben) Population mit 8 Mio. Einwohnenden, 3.6 Mio. Haushalte, 2.2 Mio. Wohnungen

Entwicklung einer «Synthetischen Bevölkerung Schweiz»

EBP hat im Rahmen eines internen Forschungsvorhabens die Methoden zur Verknüpfung und Kombination der Register- und thematischen Erhebungen zu einer synthetischen Vollerhebung entwickelt. Eine synthetische Bevölkerung erlaubt es, differenzierte, präzisere und besser belastbare Aussagen zu Bevölkerungsstruktur und Mobilitätsverhalten zu treffen, als es die Ausgangsdatensätze je für sich allein erlauben würden. Die Verknüpfungsmethoden bzw. die dadurch erzeugten Datensätze unterstützen und verbessern einerseits bestehende Beratungsdienstleitungen von EBP, wie beispielsweise die Erarbeitung von Entscheidungsgrundlagen für politische Massnahmen für die öffentliche Hand. Anderseits können sie auch direkt abgegeben werden, zum Beispiel an öffentlichen Ämtern, als Grundlage für ihre internen Analysen und Berechnungen.

Die Verknüpfungsmodelle führen zwei Echtdatensätze sowie zwei Stichprobenerhebungen zusammen zur «Synthetischen Bevölkerung Schweiz», welche als relationale Datenbank realisiert ist:

- die Gebäude- und Wohnungsstatistik (GWS 2014);

- das Motorfahrzeuginformationssystem mit allen Personenwagen (MOFIS 2015);
- die Haushaltsbudgeterhebung (HABE 2006-2011; Stichprobe);
- der Mikrozensus Mobilität und Verkehr (MZMV 2010; Stichprobe).

Für die möglichst sinnhafte Verknüpfung der Datensätze kommen weitere Datensätze (z.B. zum Steueraufkommen je Gemeinde) zum Einsatz. Weil Echtdatensätze als Kern der synthetischen Bevölkerung angewendet werden, liegt eine teil-synthetische Bevölkerung vor. Abbildung 2 zeigt eine Übersicht der Verknüpfungen, die zur Erstellung der teil-synthetischen Bevölkerung dienen.

Verknüpfungsmodelle als Hauptresultat

Die in der Programmiersprache R geschriebenen Verknüpfungsmodelle werden von EBP entwickelt und stellen die eigentliche Substanz der «Synthetischen Bevölkerung Schweiz» dar. Sie erzeugen eine synthetische Vollerhebung der HABE und des MZVM, der den PLZ-scharfen Fahrzeugbestand aus MOFIS strikt abbildet. Dabei werden die Einzeldatensätze nicht zufällig den Personen zugeteilt, sondern die Verknüpfungsmodelle berücksichtigen Haushalttyp, Mobilitäts-werkzeuge, Ausgabeverhalten, BFS-Grossregion sowie Einkommenskategorie.

Der MZVM bildet «nur» die Stichtagsmobilität einer jeden Zielperson vollständig ab. Für den Einsatz des MZVM in synthetischen Bevölkerungen genügt dies nicht, weil die Stichtagsmobilität stark abweichen kann von der Jahresmobilität einer Zielperson. Deshalb wird für jede Zielperson in der synthetischen Population überdies eine individuelle, mit dem Besitz an Mobilitätswerkzeugen kohärente MIV- und ÖV-Jahresfahrleistung berechnet. Die Gesamtheit dieser Jahresfahrleistungen wahrt die statistischen Verteilungen (gesamte Fahrleistung Schweiz; Fahrleistungen innerhalb Gruppen, die anhand mobilitätsrelevanter Variablen gebildet werden) des MZVM-Rohdatensatzes 2010.

Die «Synthetische Bevölkerung Schweiz» weist folgende Eckwerte auf:

- 8 Millionen Personen in ca. 2.2 Millionen Haushalten;
- mit einem Wohngebäude mit Bau- und Renovationsjahr sowie Energieträger für Raumwärme und Warmwasser;
- mit «Mobilitätswerkzeugen» (Autos und ÖV-Abonnemente) und zugehöriger Mobilität (Personenkilometer MIV und ÖV);
- mit einer Einkommensklasse sowie Konsumgewohnheiten (Ausgaben pro Konsumkategorie).

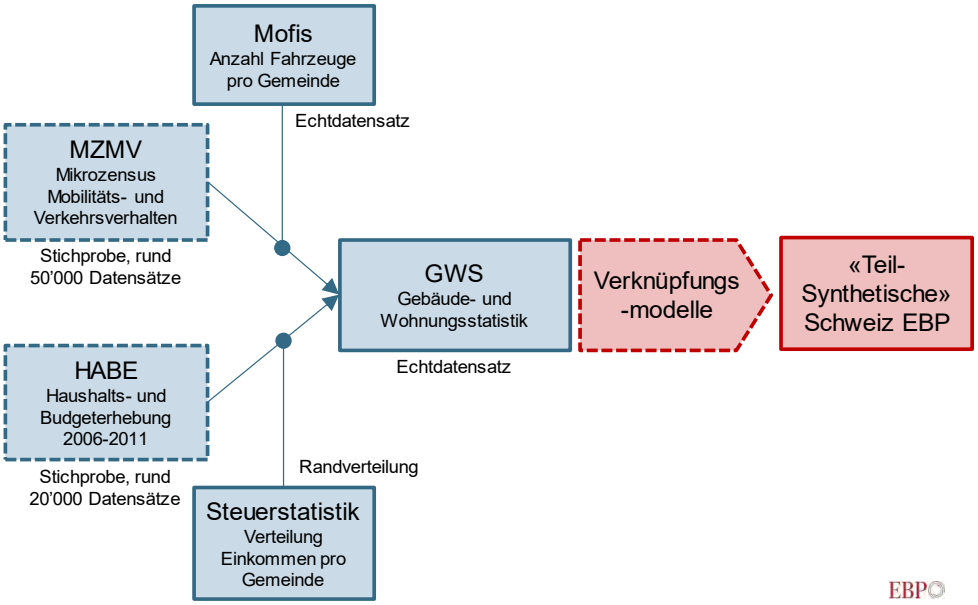


Abbildung 2: Schematische Darstellung der Teil-Synthetischen Schweiz

Prozess zur Qualitätssicherung und Abnahme mit Auftraggebern

Synthetische Bevölkerungen bieten einen grossen Mehrwert für die politische Beratung und weitere Projekte, in welchen strukturell differenzierte Aussagen (räumlich, sozio-demographisch, etc.) zu treffen sind. Die Anwendung von solchen synthetischen Datensätzen in der täglichen Beratung ist jedoch nicht trivial. Solche Projekte stellen hohe Anforderungen in der Qualitätssicherung, von der Erstellung der Datensätze bis zur Abnahme der Daten oder spezifischer Resultate durch den Auftraggeber. Unter dem Titel «Wie schweizerisch ist die synthetische Schweiz von EBP?» stellen wir diese Fragen der Qualitätssicherung und der Zusammenarbeit mit Kunden in den Vordergrund. Die Erkenntnisse beruhen auf zwei konkreten Projekten mit dem Amt für Verkehr des Kantons Zürich und den SBB:

- Im Projekt mit dem **Amt für Verkehr des Kantons Zürichs** soll der MZVM in einer synthetischen Vollerhebung abgebildet werden, so dass eine valide Datenbasis in spezifischen Regionen (z.B. Gemeinden) oder für einzelne Haushaltsgruppen vorliegt.
- In einem bestehenden Projekt mit den **SBB** stellt EBP die Nähe zur nächsten ÖV-Haltestelle für jeden Haushalt der Schweiz zur Verfügung. Mit der Synthetischen Schweiz wird geprüft, welche weiteren Informationen räumlich differenziert zur Verfügung stehen, um das Mobilitätsverhalten der Haushalte hinsichtlich ÖV besser zu verstehen.

Im Folgenden wird ein generischer Prozess beschrieben für den Umgang mit synthetischen Bevölkerungen in konkreten Beratungsprojekten und zur Qualitätssicherung in diesen Projekten. Dies geschieht anhand von Thesen, von denen im Folgenden zwei näher beschrieben werden.

These 1: Eine genaue Spezifikation der Datenlieferung zwischen Auftraggeber und Auftragnehmer ist die zentrale Grundlage.

Bei Start eines Beratungsauftrags bzw. einer Datenlieferung der synthetischen Bevölkerung ist zwischen Auftraggeber und Auftragnehmer schriftlich zu

definieren, welche Variablen (inkl. deren Ausprägungen) in welcher Form geliefert werden. Eine solche Vereinbarung legt die Grundlage für eine erfolgreiche Anwendung der Resultate bzw. Daten beim Auftraggeber.

These 2: Qualitätskriterien für die Abnahme der Daten sind vorgängig festzulegen.

Die Abnahme der Daten durch den Auftraggeber wird idealerweise vor der Datenlieferung vorbereitet, indem die Abläufe der Qualitätssicherung klar festgelegt werden. Dazu gehört insbesondere:

- Welche Qualitätskriterien werden geprüft? Für welche Variablen und auf welchem Aggregationsniveau sollen die Auswertungen der Qualitätssicherung vorgenommen werden? Die Auswahl der Prüfvariablen hat einen grossen Einfluss darauf, welche Datenquellen sich zur Qualitätssicherung eignen und in welcher strukturellen Auflösung die Qualitätssicherung vorgenommen werden kann.

- Welche Datenquellen werden verglichen und zur Qualitätssicherung herangezogen?
  Zur Qualitätssicherung dient oft ein Vergleich mit einer zweiten, (idealerweise) unabhängigen Datenquelle. Bei diesem Vergleich ist jedoch zu prüfen, ob die beiden Datenquellen tatsächlich die exakt gleiche Variable abbilden. Abbildung 3 zeigt ein solches Vorgehen anhand eines Vergleichs der Fahrleistung MIV pro Zürcher Gemeinde in 1) der synthetischen Bevölkerung EBP und 2) dem kantonalen Gesamtverkehrsmodell. Die generellen Zusammenhänge werden gut erklärt, jedoch ist die Fahrleistung im Gesamtverkehrsmodell systematisch höher. Diese Abweichung ist auf unterschiedliche Definitionen zurück zu führen.

- Welche Qualitätsgüte soll für welche strukturelle Auflösung erreicht werden? Beispielsweise könnten einige Variablen robust genug sein, damit sie auf Ebene Hektaren ausgewertet werden können. Andere Variablen dürften nur für Gemeinden (mit einer minimalen Anzahl Einwohnern) ausgewertet werden. Diese Vorgaben sind idealerweise detailliert zu spezifizieren.

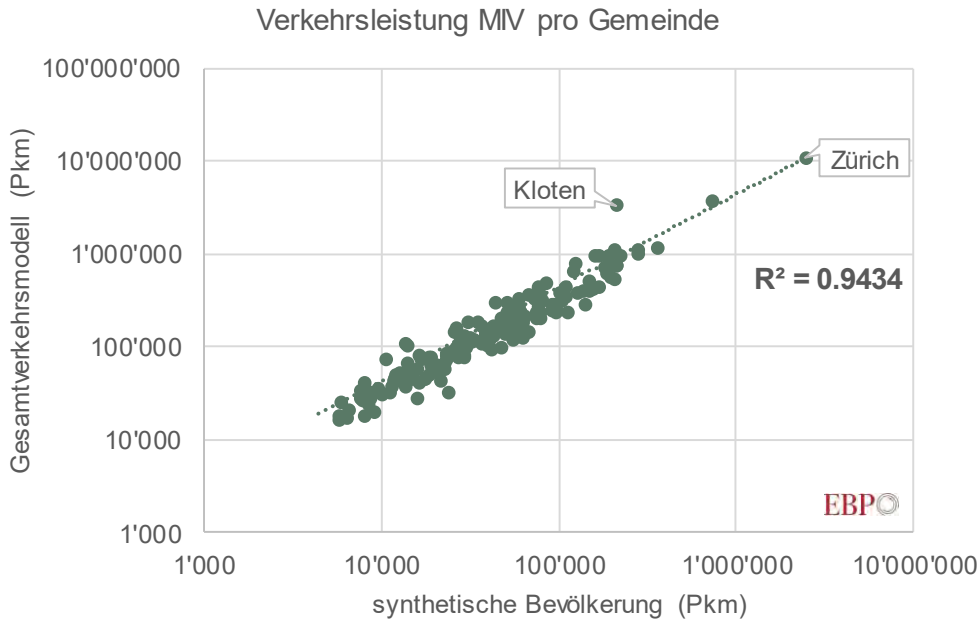


Abbildung 3: Vergleich der synthetischen Bevölkerung mit dem Gesamtverkehrsmodell Kanton Zürich zur Qualitätssicherung. Verglichen wird die Kenngrösse Fahrleistung MIV pro Gemeinde.

2.3 Aufbau und Anwendung einer synthetischen Population im Verkehrsmodell Oberösterreich (Haupt, th-inc / Senozon)

Synthetische Population für Oberösterreich	
Ersteller	Planoptimo, Büro Dr. Köll ZT GmbH aus A-Reith bei Seefeld
Partner	th-inc GmbH, Thomas Haupt aus D-Karlsruhe und Senozon AG, Zürich.
Software	EXCEL, ACCESS, VISUM (Alle proprietär)
Themengebiete	Verkehrsmodellierung
Zeitliche & Geographische Auflösung	Stichtag im Oktober 2012; adressscharfe Auflösung, anonymisierte Haushaltsbefragung mit 80000 Haushalten, ca. 370000 Adresspunkte
Nächste Schritte	Bereits in ein landesweites Verkehrsmodell eingebaut
Prognosefähigkeit	Ja, in separaten Zeitscheiben, Prognose zunächst nur auf Verkehrszellenebene
Grösse der grundlegenden Datenmenge	SynPop für 1,4 Mio Einw

SynPop in einem makroskopischen Verkehrsmodell

Beim Aufbau des Verkehrsmodells Oberösterreich konnten neue Wege beschritten werden, da erstens eine aktuelle, äußerst umfangreiche Haushaltsbefragung vorlag und zweitens sehr umfangreiche, adressfeine Primärdaten sowie umfangreiche weitere Daten.

In einer prototypischen Implementierung wurden vorab grundsätzliche Fragen geklärt:

- Können 1,4 Mio Personen als Point of Interests in einem Verkehrsmodell verwaltet werden?
- Können durch Verschneidung innerhalb von Verkehrszellen alle Strukturmerkmale und aggregierte verhaltenshomogene Personengruppen widerspruchfrei zur bestehenden Statistik gewonnen werden?
- Kann durch die Nutzung der synthetischen Population die Genauigkeit der Modellierung verbessert werden?
- Können alle Verfahrensschritte in einem E-xtract, T-ransform, L-oad Prozess automatisiert und wiederholbar in dem Verkehrsmodell implementiert werden?
- Reichen die Rechenkapazitäten und die Verarbeitungsgeschwindigkeit aus?

Statistische Gliederung der Eingangsdaten

Das räumliche Bezugsgebiet ist das Bundesland Oberösterreich mit seiner verwaltungstechnischen Gliederung in 444 Gemeinden, 3 Statutarstädte (Linz, Wels und Steyr) und 15 politische Bezirke.

Die räumliche Feinaufteilung der Gemeinden umfasst 1.266 sogenannte Zählsprengel, welche die kleinsten Einheiten darstellen, für die statistische Datengrundlagen (Wohnbevölkerung, Haushalte, Arbeitsplätze, etc.) verfügbar sind. Diese Zählsprengel entsprechen gleichzeitig den im Verkehrsmodell abgebildeten Verkehrszellen von Oberösterreich.

In der Haushaltserhebung gibt es neben der adressscharfen Codierung der Wohnadresse eine Zuordnung zu VE-Gebieten, die einer von 444 Gemeinden und in den größeren Städten einer Aggregation von Zählsprengeln zu Stadtteilen entspricht.

Für weitere Auswertungen der Haushaltsbefragung und zur Visualisierung im Verkehrsmodell wurden die Adresskoordinaten randomisiert, so dass jede X/Y Koordinate einen Einzugsbereich von mindestens 5 Haushalten hat.

Des Weiteren lag ein 250m-Raster mit den aus der Realbevölkerung gefüllten Bevölkerungszahlen vor (Hauptwohnsitzmeldungen ohne Zusatzinformation), das ca. 56.000 Raster umfasste, sowie Shape-Dateien der Flächennutzung, die eine Abgrenzung der bebauten Gebiete ermöglichte.

Vorgehen bei der Bildung der SynPop

Wesentliche Grundlage für die Erzeugung der synthetischen Bevölkerung war die Klassifikation der Gesamtbevölkerung nach 6 Merkmalen (Zählsprengel, Haushaltstyp, Alter, Geschlecht, Erwerbsstatus, Erwerbsart) durch die Statistikabteilung des Landes Oberösterreich. Bemerkenswert daran ist, dass als Ergebnis eine mehrdimensionale Kreuztabelle vorliegt, welche die tatsächliche **gemeinsame** Verteilung der genannten Eigenschaften in der Grundgesamtheit wiedergibt, und gleichzeitig - trotz der aus Datenschutzgründen erfolgten 'Verwischungen' (Target Swapping) in dünn besetzten Zellen - die Konsistenz mit den **Randverteilungen** gewahrt bleibt. Zu den Einträgen in der Kreuztabelle wurden mittels EXCEL-Makro Datensätze generiert, welche neben der Person(ennummer) auch die bekannten Attribute enthalten. Weitere für die vorgesehene Verwendung im Verkehrsmodell relevante Eigenschaften (Pkw-Verfügbarkeit, Teil-/Vollzeitbeschäftigung, Zugehörigkeit zur Gruppe der Lehrlinge/Azubis) mussten anschließend zugespielt werden. Informationen dazu lieferte die Haushaltsbefragung. Zu berücksichtigen waren insbesondere die Korrelationen der Zusatzmerkmale untereinander und zu den übrigen Personeneigenschaften wie Alter und Geschlecht.

Im hier beschriebenen Anwendungsfall wurde auf der Haushaltsebene lediglich zwischen Anstalts- und Privathaushalten unterschieden und nur im erstgenannten Fall, das heisst bei Wohn-, Lehrlings- und Studentenheimen sowie Straf- und Justizvollzugsanstalten synthetische Personen auf synthetische (Anstalts-) Haushalte aufgeteilt. Bei den in Privathaushalten lebenden Personen unterblieb eine solche Zuordnung auf synthetische (Privat-)Haushalte gegebener Größe und Charakteristik, weil der dazu erforderliche Mehraufwand größer eingeschätzt wurde als der erzielbare zusätzliche Nutzen.

Die Übernahme der synthetischen Bevölkerung ins Verkehrsmodell, im Wesentlichen also deren Verortung, erfolgte wiederum getrennt für Anstaltshaushalte und Personen in Privathaushalten. Bei den rund 200 größeren Anstaltshaushalten des Landes (hauptsächlich Wohn- und Pflegeheime für SeniorInnen) sind die Adressen bekannt, BewohnerInnen kleinerer Einrichtungen wurden anteilig auf größere aufgeteilt. Für die übrigen Personen (rund 98% der gesamten Wohnbevölkerung) wurden zuerst die Einwohnerzahlen pro Rasterquadrat bereinigt um Personen in Anstaltshaushalten und solche, die allenfalls außerhalb des Bezugsgebietes liegen (für den Fall, dass eine Landesgrenze durch ein Rasterquadrat verläuft). Anschließend wurden die Adresspunkte in jedem Rasterquadrat mit Hilfe der bekannten Flächenwidmung qualifiziert (Wohnnutzung vs. sonstige) und auf Basis der Verschneidung von Zählsprengeln und Rasterquadraten in einem eigens dafür entwickelten und in VISUM implementierten Algorithmus so gewichtet, dass die Sollwerte für die synthetischen Personen auf Raster- und Zählsprengelebene möglichst exakt erreicht werden.

Abbildung 4: Räumliche Verteilung der synthetischen Bevölkerung auf Adressen mit Wohnnutzung (rot), und Anstaltshaushalte (gelb) innerhalb von Zählsprengeln (gelb) und Rasterquadraten (türkis)

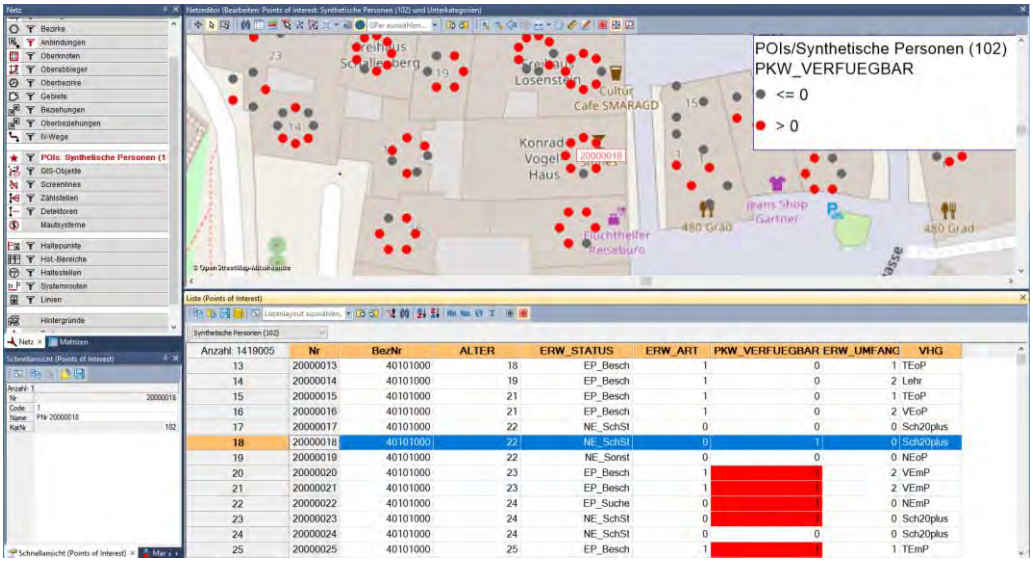
Aus Kombination von Adressgewichten und -koordinaten mit den Datensätzen der synthetischen Bevölkerung entsteht eine ideale Datenbasis, die sowohl Auswertungen mit Hilfe relationaler Datenbanksysteme (im konkreten Fall MS-ACCESS) als auch räumliche Verschneidungen und Visualisierungen (im Verkehrsmodell) ermöglichen.

Hybride Nutzung der SynPop im Verkehrsmodell

Im makroskopischen Verkehrsmodell dient die synthetische Bevölkerung der Prüfung der Datenkonsistenz und präzisen Generierung von Anbindungen mit Einwohnergewichten für Verkehrszwecke und Verkehrsmittel. Insgesamt werden

18 verhaltenshomogene Gruppen in ca. 1200 Verkehrszellen automatisiert durch Verschneideoperationen gebildet.

Für den künftigen Einsatz einer agentenbasierten Modellierung wurden mit der SynPop die datentechnischen Voraussetzungen geschaffen, da jeder synthetischen Person bereits alle für die disaggregierte Modellierung notwendigen individuellen Merkmalen zugeordnet wurden.



	Nr	BezNr	ALTER	ERW_STATUS	ERW_ART	PKW_VERFUEGBAR	ERW_UMFANG	VHG
13	20000013	40101000	18	EP_Besch	1	0	1	TEoP
14	20000014	40101000	19	EP_Besch	1	0	2	Lehr
15	20000015	40101000	21	EP_Besch	1	0	1	TEoP
16	20000016	40101000	21	EP_Besch	1	0	2	VEoP
17	20000017	40101000	22	NE_SchSt	0	0	0	Sch20plus
18	20000018	40101000	22	NE_SchSt	0	1	0	Sch20plus
19	20000019	40101000	22	NE_Sonst	0	0	0	NEoP
20	20000020	40101000	23	EP_Besch	1	1	2	VEmP
21	20000021	40101000	23	EP_Besch	1	1	2	VEmP
22	20000022	40101000	24	EP_Suche	0	1	0	NEmP
23	20000023	40101000	24	NE_SchSt	0	1	0	Sch20plus
24	20000024	40101000	24	NE_SchSt	0	0	0	Sch20plus
25	20000025	40101000	25	EP_Besch	1	1	1	TEmP

Abbildung 5: Einwohnerliste mit Verortung im Verkehrsmodell Oberösterreich

In der Anwendung zeigte sich, dass die Datenmengen im Verkehrsmodell sehr gut verarbeitet werden können und dem Verkehrsplaner und Modellierer mit der punktgenauen Verortung einerseits ein hochaufgelöstes System bereitgestellt werden kann, gleichzeitig aber auch alle (räumlichen und soziodemografischen) Aggregationen und rechenzeitoptimierte Verfahren eines makroskopischen Verkehrsmodells zur Verfügung stehen. Ausserdem wurden damit die Voraussetzungen geschaffen künftig – voraussichtlich mit MATSim (Horni et al. 2016) – ein agentenbasiertes Modell zu betreiben.

2.4 Generierung synthetischer Bevölkerungen für Berlin - Möglichkeiten und Grenzen (Cyganski, DLR)

Name:	Antworten
Ersteller	Institut für Verkehrsforschung, Deutsches Zentrum für Luft- und Raumfahrt e.V.
Partner	-
Software	SYNTHESIZER (Freigabe open source geplant)
Themengebiete	Bisher v.a. Verkehrsnachfragemodellierung; (Wohn-) Standortwahlmodellierung
Zeitliche & Geographische Auflösung	Referenzjahr 2010, Grossregion Berlin, Haushalte mit Adresskoordinaten für jede Jahresscheibe
Nächste Schritte	Nutzung für weitere Szenarienrechnungen der Verkehrsnachfrage und Standortwahl; Fortschreibung Strukturgrößen über ein Evolutionsmodell; Neuberechnung Regressionsmodelle für die Mobilitätswerkzeuge mit aktualisierten Datenquellen
Prognosefähigkeit	Ja, separate Zeitscheiben für 2020, 2030, 2040
Grösse der grundlegenden Datenmenge	Stichprobengröße variierend je nach Untersuchungsraum; in der Regel aufbauend auf Mikrozensus (0.1% aller Haushalte)

Anforderungen an die synthetische Population

Künstlich generierte Bevölkerungen gehören zu den wichtigsten Eingangsdaten von Verkehrs- und Standortwahlmodellen. Güte und Genauigkeit der mit den Simulationen erzeugten Ergebnisse hängen dabei direkt mit der Qualität der genutzten Eingangsdaten zusammen. Die konkreten Anforderungen an die Genauigkeit und die Attribute der synthetischen Bevölkerungsdaten sind dabei abhängig von der Art des Modelleinsatzes und der avisierten Fragestellung. Am Institut für Verkehrsforschung des Deutschen Zentrums für Luft- und Raumfahrt e.V. (IVF) finden entsprechende Eingangsdaten bisher ihre Anwendung für die Berechnung der Verkehrsnachfrage im mikroskopischen Nachfragemodell TAPAS (Heinrichs et al. 2016) sowie dem auf dem Simulationspaket Cube Land basierenden Standortwahlmodell SALSA (Martinez et al 2010; Heldt et al. 2017). Im Fokus liegt dabei in der Regel der Untersuchungsraum Berlin, für den je nach Projektanforderung verschiedene Zeitscheiben sowie Prognosen der Bevölkerungsentwicklung genutzt werden.

Für eine automatisierte Erstellung der synthetischen Bevölkerung wurde eine modulare, Java-basierte Applikation entwickelt, die die Nutzung verschiedener gängiger Verfahren zu Erstellung einer Bevölkerung auf Basis einer vorliegenden Stichprobe ermöglicht. Der sogenannte SYNTHESIZER (von Schmidt et al. 2017) ermöglicht a) die einfache Hochrechnung einer Stichprobe anhand eines gegebenen Hochrechnungsfaktors, b) die Anpassung einer Stichprobe an

Haushalts- oder Personenrandsummen oder c) die gleichzeitige Einhaltung gegebener Haushalts- und Personenrandsummen (vgl. Abbildung 6).

Variante A stellt die einfachste Art dar, auf Basis einer gegebenen, repräsentativen Stichprobe aus der Gesamtbevölkerung sowie eines bekannten Hochrechnungsfaktors eine synthetische Bevölkerung des Untersuchungsgebietes zu erzeugen. Hierbei wird jeder Haushalt samt der dazugehörigen Personen so oft aus der Stichprobe kopiert, wie er nach dem jeweiligen Hochrechnungsfaktor vorkommt (vgl. Moekel 2016).

Auch *Variante B* startet mit einer disaggregierten Bevölkerungsstichprobe. Darüber hinaus setzt sie das Vorhandensein von Randsummen all derjenigen Attribute voraus, hinsichtlich derer die Zielbevölkerung eine korrekte Verteilung aufweisen soll. Dabei kann es sich entweder um personen- oder haushaltsbezogene Randsummen handeln. Mit Hilfe mathematischer Anpassungsverfahren, in der Regel des sogenannten Iterative Proportional Fitting (IPF) (siehe Beckman et al. 1996; Farooq et al. 2013), werden im Laufe der Erstellung der synthetischen Bevölkerungen die Wahrscheinlichkeiten für die Auswahl eines Haushaltes oder einer Person derart iterativ angepasst, dass die Randsummen getroffen werden. Das Verfahren läuft solange, bis eine vorab definierte Qualitätsschwelle oder maximale Iterationszahl erreicht wurde.

Variante C erweitert das Vorgehen aus Variante B dahingehend, dass gleichzeitig vorliegende Randsummen für die Attribute der Personen sowie der Haushalte eingehalten werden. Neben dem IPF-Verfahren kommt hier in der Regel das sogenannte Iterative Proportional Updating (IPU; siehe von Schmidt et al. 2017, Müller und Axhausen 2011 sowie Ye et al. 2009) Verfahren zur Anpassung der Haushaltsgewichte zum Einsatz.

Detaillierte Informationen zum Vorgehen sowie den Vor- und Nachteilen der einzelnen Varianten finden sich in (von Schmidt et al. 2017); Müller und Axhausen (2011) gibt einen Überblick über unterschiedliche Varianten und Erweiterungen des IPF. Eine Vorstellung des IPU findet sich bei Ye et al. (2009). In der Regel findet dabei bei der Erstellung von synthetischen Bevölkerungen für die Simulationsmodelle am Institut für Verkehrsforschung die Variante C Verwendung, bei der die Einhaltung der vorgegebenen Randsummen mit Hilfe iterativer IPF- und IPU-Anpassungen erfolgt.

Wie anhand der Abbildung 6 ersichtlich, erfolgt die Erstellung der synthetischen Bevölkerung in drei Arbeitsschritten, die sich im Vorgehen sowie der zugrundeliegenden Datenbasis stark unterscheiden. Zunächst wird die Grundbevölkerung mit Hilfe eines der oben genannten Verfahren erstellt. Die nachfolgend näher beschriebenen Datengrundlagen weisen dabei zwei Schwachpunkte auf, die in den beiden anschließenden Bearbeitungsschritten kompensiert werden sollen. Zunächst liegen die Randsummen meist mit einer geringen räumlichen Auflösung vor, sodass anschließend eine räumliche Vereinzelung mit der Zuweisung einer konkreten X-Y-Koordinate erfolgt. Zudem weisen die verwendeten Daten keine mobilitätsbezogenen Informationen auf. Angaben zum Besitz eines Führerscheins, eines ÖV-Abos und eines Pkws werden daher mit Hilfe spezifischer Regressionsmodelle in einem weiteren Arbeitsschritt angespielt. Je nach Projektkontext kann hier auch eine Differenzierung der Pkw-Flotte vorgenommen werden (vgl. z.B. Heinrichs et al. 2016). Für alle drei Bearbeitungsschritte sollen die verwendeten Datenquellen, Attribute und Einschränkungen nachfolgend kurz adressiert werden.

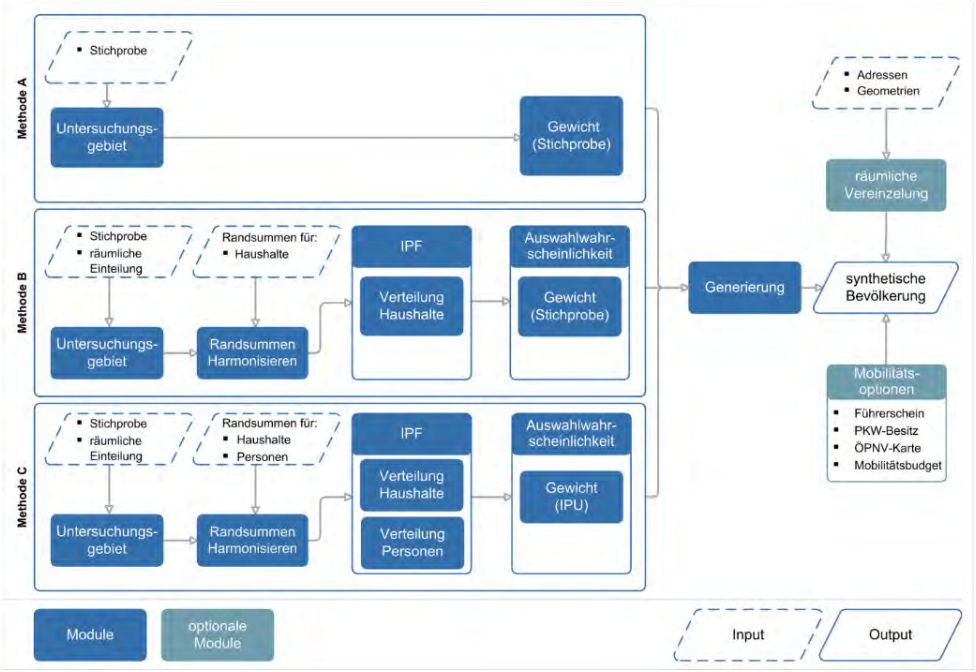


Abbildung 6: Methoden der Erstellung synthetischer Bevölkerungen mit dem SYNTHESIZER.

Für die Erstellung der Basisbevölkerung wird auf den deutschen Mikrozensus (Statistisches Bundesamt 2018) als Stichprobenbasis zurückgegriffen. Er enthält umfassende Informationen zu den einzelnen Personen, ihrem Haushaltskontext und ihrer Wohnsituation. Insgesamt beinhaltet er über 100 Attribute, von denen jedoch nur eine Auswahl bei der Bevölkerungsgenerierung hinsichtlich ihrer Verteilung berücksichtigt wird. Detaillierte Angaben sind hier nur auf Bundeslandebene zu erhalten, Randsummen auch auf Bezirksebene. Für die Erstellung der Randsummen, hinsichtlich derer die Bevölkerung im Erstellungsverfahren geprüft wird, wird für die Basisbevölkerung in der Regel ebenfalls auf den Mikrozensus zurückgegriffen. So vorhanden können hier jedoch auch räumlich spezifischere Daten Verwendung finden: im Falle des Untersuchungsraumes Berlin liegen so teilweise Daten der Berliner Senatsverwaltung auf Ebene der Berliner Bezirke vor. Insgesamt lässt sich konstatieren, dass die benötigten Daten zur Bestimmung der Randsummen bereits für die Abbildung des Basiszustandes nur teilweise in der benötigten Art und räumlich differenzierten Auflösung zur Verfügung stehen und oftmals aus unterschiedlichen Quellen stammen. Dies gilt insbesondere für die Informationen zum Berufsstatus der Personen sowie zum Haushaltseinkommen. Dies führt in der Regel zu aufwändigen Arbeiten für die Harmonisierung der Randsummen sowie zur Notwendigkeit der Festlegung der wichtigsten Referenzquelle.

Zu den häufig notwendigen Harmonisierungsschritten zählen:

- Umwandlung von Attributausprägungen (z.B. Zusammenfassen bzw. Aufteilen von Altersgruppen);
- räumliche Randsummenanpassung (z.B. ältere Verteilung auf Teilverkehrszellenebene an neue Daten auf Bezirksebene anpassen);

- Randsummenanpassung innerhalb einer Ebene (z.B. Gesamtanzahl der Haushalte nach HH-Einkommen anteilmäßig an die Gesamtanzahl der Haushalte nach HH-Größe anpassen);
- Randsummenanpassung zwischen den Ebenen (z.B. Gesamtanzahl der Haushalte auf HH-Größe anteilmäßig an die Gesamtanzahl der Personen anpassen).

Datengrundlagen

Als wichtigste Datenquellen für die Basissituation sind der Mikrozensus zu nennen, für die Prognosezeitpunkte wird in der Regel auf die Bevölkerungsfortschreibung des Bundesamtes für Bau-, Stadt- und Raumforschung (BBSR 2012) zurückgegriffen.

Für die präzise räumliche Verortung der so generierten Grundbevölkerung auf Ebene der Bezirke werden Daten des Digitalen Landschaftsmodells (DLM) des Bundesamts für Kartographie und Geodäsie (BKG), des amtlichen Liegenschafts-katasterinformationssystems (ALKIS) sowie der Adressdatensatz des Bundesamts für Kartographie und Geodäsie (BKG) verschnitten. Auf diese Weise wird ermittelt, welche Adresskoordinaten für die Zuweisung eines Wohnstandortes genutzt werden sollten. Die Grundbevölkerung wird dann entsprechend in den einzelnen Häusern positioniert, und die Koordinateninformationen können für die Generierung einer hochauflösenden Nachfrage eingesetzt werden.

Angaben zu den Mobilitätsoptionen liegen in der Regel nicht an den für die Generierung der Grundbevölkerung genutzten Daten vor. Auf Basis von spezifischen Verkehrserhebungen, für Berlin dem Datensatz der Erhebung Mobilität in Städten (SrV) (TU Dresden 2018), werden separate Regressionsmodelle zur Ausstattung der Personen und Haushalte mit Führerscheinen, ÖV-Karten sowie Pkws geschätzt. Anhand der Attribute der synthetischen Bevölkerung werden anschließend die jeweiligen Ausprägungen an die Basisbevölkerung angespielt.

Tabelle 1 zeigt eine Übersicht der Attribute und ihrer Ausprägungen der fertigen synthetischen Bevölkerung. Fett dargestellt sind dabei diejenigen Attribute, die bei der Erstellung der Bevölkerung mit Hilfe der Methode Variante C hinsichtlich ihrer Randsummen geprüft werden.

Attribut	Attributausprägung
HH-ID	…
HH-Größe	1, 2, 3, 4 und 5+ Personen
HH-Einkommensgruppe (Euro)	0 - 899, 900 - 1499, 1500 - 1999, 2000 - 2599, 2600 - 3199, 3200+
HH-Einkommen (Euro)	Für jede Einkommensgruppe wird eine gleichverteilte Zufallszahl erzeugt, um somit diskrete Einkommenswerte zu ermitteln. Bei Prognosejahren wird zusätzlich noch eine jährliche Wachstumsrate hinzuaddiert.
Kinder im HH	Ja/Nein
HH-Typ	TAPAS-HH-Typ z.B. Zweipersonenhaushalt (2 Erwachsene), Zweipersonenhaushalt (1 Erwachsener mit Kind)
Anzahl Autos im HH	0/1/2
Welches bzw. welche Autos?	z.B. Größe/Antrieb/Automatisierung
Verkehrszelle-ID	z.B. Bezirk/TVZ
Koordinaten	Adresse
Personen-ID	…
HH-ID	…
Geschlecht	männlich, weiblich
Alter	0, 1, 2, …, 100
Status	Nichterwerbspersonen (NEP) Kind unter 6, Schüler, Student, Rentner, sonstige NEP Erwerbspersonen (EP) Vollzeit, Teilzeit, Erwerbslos
Personen-Typ	TAPAS-Personen-Typ z.B. erwerbstätig, Mann oder Frau, kein Pkw im HH, bis 24 Jahre
Führerschein	JA/NEIN
ÖPNV-Ticket	JA/NEIN
Fahrrad	JA/NEIN
MB-MIV-Variable	Mobilitäts-Budget: MIV-Variable (z.B. für Kraftstoff, Wartung)
MB-MIV-Fix	Mobilitäts-Budget: MIV-Fix (z.B. für Steuern, Versicherung, Kfz-Erwerb)
MB- ÖPNV	Mobilitäts-Budget: ÖPNV (z.B. für Monatsticket)

Tabelle 1: Übersicht der Bevölkerungsattribute und ihrer Ausprägungen; bei der Erstellung nach Variante C mit Hilfe der Randsummen kontrollierte Attributsverteilungen sind fett markiert

Weiterführende Arbeiten: Prognose, Qualitätskontrolle und Ausdehnung des Untersuchungsraumes

Die Erstellung der synthetischen Bevölkerung erfolgt momentan für jede betrachtete Zeitscheibe separat; eine Fortschreibung über die Jahre mit Hilfe von Evolutionsmodelle etc. ist bisher noch nicht umgesetzt. Eine Aktualisierung der Bevölkerung wird in Abhängigkeit von den jeweiligen Anforderungen der Projekte vorgenommen, die oftmals unterschiedliche Anforderungen hinsichtlich der Datenaktualität, der Konsistenz mit anderen Eingangsdaten oder spezifischen räumlichen Bezügen aufweisen.

Aufgrund der unterschiedlichen Datenquellen kommen für die Prüfung der Qualität der erstellten Bevölkerung unterschiedliche Verfahren zum Einsatz. Für die erstellte Grundbevölkerung findet eine Prüfung der Abweichung der Attributsverteilungen in der erstellten Bevölkerung von den bekannten Randsummen statt. Als Indikator wird hier der oft verwendete Standardized *Root Mean Square Error (SRMSE)* berechnet (Prichard und Miller 2012). Da die Randsummen oftmals nur auf hoher Aggregationsebene vorliegen, finden darüber hinaus kartenbasierte visuelle Prüfungen der absoluten und relativen Bevölkerungsanteile und ihre Veränderung zwischen verschiedenen Zeitscheiben Anwendung. Diese werden auch genutzt, um die feinräumliche Verteilung der Bevölkerung nach der Vereinzelung zu prüfen. Die Prüfung der mobilitätsbezogenen Eigenschaften der Bevölkerung findet im Abgleich mit den für die Regressionsmodelle verwendeten Eingangsdaten, in der Regel der SrV (TU Dresden 2018) oder der MiD (BMVI 2018), statt. Hier werden die jeweiligen Anteile für die vorliegenden Bezugseinheiten berechnet sowie eine kleinräumigere Prüfung anhand von Karten vorgenommen. Eine Prüfung der Verteilungen der im Mikrozensus vorhandenen zusätzlichen Attribute, die nicht bei der Erstellung der Bevölkerung als Randsummen Eingang finden, wird nicht vorgenommen.

Der SYNTHESIZER wurde bisher vorrangig zur Erstellung der Bevölkerungen für Berlin sowie den Untersuchungsraum Braunschweig eingesetzt. Der flexible, modulare Aufbau ermöglicht jedoch eine schnelle Anpassung auch für andere Einsatzzwecke. Neben einer potentiellen Nutzung für die Generierung passender Eingangsdaten für das deutschlandweite makroskopische Personenverkehrsnachfragemodell des IVF wird momentan der Einsatz für die Fortschreibung von Strukturdaten für die Nachfragemodellierung geprüft. Auf Basis einer vorliegenden Vollerhebung von Standorten des Lebensmitteleinzelhandels im Jahr 2015 erfolgte eine nach Art und Größenklasse differenzierte Fortschreibung der Einkaufsorte 2030 mit Hilfe der Methode Variante B.

2.5 Simulation-based population synthesis using Gibbs sampling – the Brussels case (Bierlaire, EPFL)

Microsimulations of urban systems require a synthetic population as a key input. This input suggests to use Gibbs sampling rather than Iterative Proportional Fitting (IPF) and Combinatorial Optimization based techniques. The key shortcomings of the commonly used procedures to construct synthetic populations include:

a) fitting of one contingency table, while there may be other solutions matching the available data;

b) due to cloning rather than true synthesis of the population, losing the heterogeneity that exists in the real population but may not have been captured in the available microdata;

c) over reliance on the accuracy of the data to determine the cloning weights;

d) poor scalability with respect to the increase in number of attributes of the synthesized agents.

In order to overcome these shortcomings, the Gibbs sampling is put forward. Gibbs sampling is a Markov chain Monte Carlo (MCMC) algorithm for obtaining a sequence of observations which are approximated from a specified multivariate probability distribution, when direct sampling is difficult. Partial views of the joint distribution of agent's attributes that are available from various data sources can be used to simulate draws from the original distribution. The real population from the Swiss census was used to compare the performance of simulation based synthesis with the standard IPF. Tabelle 2 shows standard root mean square error (SRMSE) statistics which indicate that even the worst case simulation based synthesis with MCMC (SRMSE = 0.35) outperformed the best case IPF synthesis (SRMSE=0.64).

Input	IPF	Simulation
20% Sample	0.637	–
10% Sample	0.708	–
5% Sample	0.750	–
3% Sample	0.910	–
1% Sample	1.420	–
FullCond	–	0.130
Partial_1	–	0.240
Partial_2	–	0.340
Partial_3	–	0.350

Tabelle 2: Standard Root Mean Square Error (SRMSE) of two different methods to build an agent population

This methodology was also used to generate the synthetic population for Brussels, where the data availability was highly limited (0.1% or 1367 households). When the size of a data set is as small as it was for Brussels, parametric models have to be used to construct the conditional distributions based on estimates from various sources. The simulated population reproduced the observed income distribution with a reasonable fit (Abbildung ; R^2 = 0.799).

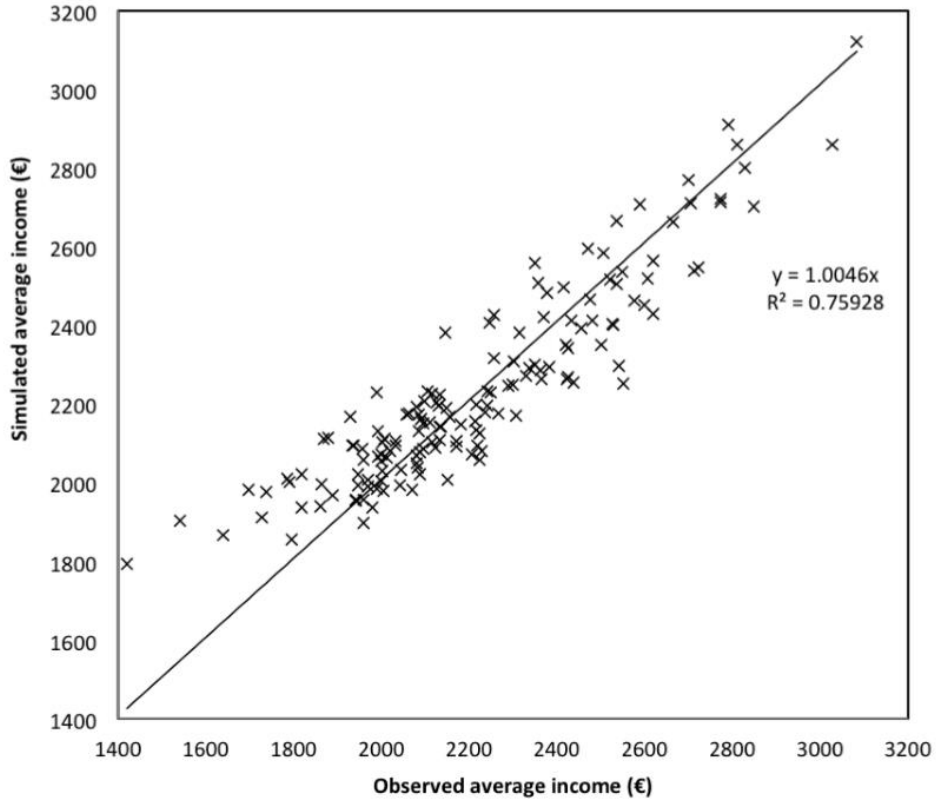


Abbildung 7. Correlation between observed income and income simulated with MCMC

3. Anwendungen

3.1 Qualitätssicherung bei synthetischen Bevölkerungen (Moser, Stat. Amt ZH)

Voraussetzungen

Der Qualitätssicherung kommt bei der Erarbeitung synthetischer Populationen eine entscheidende Rolle zu: Denn schliesslich geht es darum, eine nur teilweise bekannte Gesamtpopulation, welche die Synthetisierung überhaupt erst erforderlich macht, zu modellieren, was Wissen über die in ersterer herrschenden Zusammenhänge voraussetzt. Das Statistische Amt des Kantons Zürich begleitete deshalb auf Anstoss des kantonalzürcherischen Hauptstakeholders, des Amts für Verkehr (AfV), während rund eines Jahres die Erarbeitung der synthetischen Population für die Schweiz von EBP (Kapitel 2.2).

Eine derartige kritische Begleitung durch unabhängige Experten mit vertieften Fachkenntnissen zu den verwendeten Datensätzen aber auch sozialwissenschaftlichem Synthesewissen liefert wertvolle Einsichten für die Qualitätskontrolle. So zeigte sich beispielsweise, dass die Schätzung einer individuellen Jahresmobilität auf der Grundlage der Angaben zur Stichtagsmobilität und des ÖV-Abonnementsbesitzes im MZMV für den öffentlichen Verkehr übers ganze Jahr gesehen eine stark überhöhte Durchschnittskilometerleistung zur Folge hatte. Der Hinweis des Fachexperten des Amtes auf eine bekannte Schwäche des MZMV (die befragungstechnisch verursachte Überschätzung der Verbreitung des Besitzes von Generalabonnementen in der Bevölkerung) trug zur Verbesserung der Qualität der SynPop bei. Auch ein Vergleich der räumlichen Verteilung der Einkommen - für die synthetische Population der Haushaltsbudgeterhebung (HABE) des BFS entnommen – mit steuerstatistischen Vergleichsgrössen lieferte Einsichten, welche eine Modifikation des Zuordnungsalgorithmus nahelegten.

Erkenntnisse

Aus diesen Erfahrungen lassen sich auch einige generalisierbare Erkenntnisse gewinnen. Zunächst ist es zweifellos wichtig, dass die Reproduktion der Randverteilungen im Imputationsprozess auf dem Aggregationsniveau auf dem sie zwingend gefordert werden muss, ständig überprüft wird. Wie das obige Beispiel illustriert, müssen jedoch fallweise auch die Abgrenzung der Inputdatensätze hinterfragt werden.

Die Korrektheit in diesem technischen Sinne stellt allerdings noch nicht sicher, dass eine SynPop auch einen echten Mehrwert bringt. Die modellierten Individuen und Haushalte sollen ja – unter Bewahrung der in sozialen Kontexten stets erheblichen Variabilität – plausible Eigenschaftenbündel in den modellierten Sachbereichen aufweisen: im vorliegenden Fall sollten Wohnort, Wohnungsgrösse, finanzielle Ressourcen, die Verfügbarkeit von Mobilitätswerkzeugen und das Mobilitätsverhalten in Ausmass und Vorzeichen „wirklichkeitsgetreu“ zusammenhängen. Dafür ein einfaches Testprotokoll mit klaren Soll-Ist Toleranzen zu formulieren ist freilich kaum möglich und auch nicht sinnvoll. Dennoch sollte auch dieser anspruchsvolle, Kreativität und Sachwissen erfordernde Aspekt der Qualitätssicherung vorgängig strukturiert werden. So kann etwa das räumliche Aggregationsniveau, auf dem sinnvolle Zusammenhänge erwartet werden können, im Voraus definiert werden, ebenso wie die „stilisierten

Fakten“, die grossen gesellschaftlichen Strukturzusammenhänge, welche die synthetische Bevölkerung abbilden soll.

Nicht zuletzt erfordert der kritische Blick des potentiellen Nutzers auch einen adäquaten Zugriff auf das Endresultat, den Datensatz der SynPop. Die Werkzeuge und die Kompetenzen zur Bearbeitung grosser Datensätze stehen nicht jedem zur Verfügung. Auch diesbezüglich sollte der Nutzer vorgängig spezifizieren, in welcher Form er auf den Datensatz zugreifen will oder kann. Eine Möglichkeit den Zugriff zu vereinfachen bestünde etwa darin, das Endprodukt durch eine intuitive Benutzeroberfläche zu vermitteln, die heute ja mit geringem Programmieraufwand (z.B. in R-shiny) erstellt werden kann.

3.2 Bedürfnisse der Bundesverwaltung & Einsatz in den Themen Raum und Verkehr (Justen, ARE)

Das Bundesamt für Raumentwicklung ARE ist seit 2011 an der Entwicklung eines Flächennutzungsmodells welches die Prognose von Bevölkerung und Arbeitsplätzen erlaubt beteiligt (Facility Location Choice Simulation, FaLC, siehe ARE, 2014; ARE, 2017; und Kapitel 2.1). Das ARE verfolgt dabei das Ziel Planungen und Entscheidungsprozesse im Kontext der Raumentwicklung mittels quantitativer Analysen zu unterstützen. Zentrale und z.T. wiederkehrende Fragestellungen betreffen dabei die zukünftige, kleinräumige Verteilung von Bevölkerung und Arbeitsplätzen sowie Fragen mit Bezug zur Raumordnungspolitik. Auch sollen notwendige Eingangsdaten für die Verkehrsmodellierung (z.B. Bevölkerung nach sozioökonomischen Merkmalen, Arbeitsplätze nach Branchen) durch die Flächennutzungsmodellierung bereitgestellt werden. In Zukunft strebt das ARE die Etablierung eines integrierten Modells der Verkehrs- und Flächennutzungsmodellierung an. Ziel ist es, mit der Erarbeitung der nächsten Verkehrsperspektiven erstmals ein vollständig gekoppeltes Verkehrs- und Flächennutzungsmodell für die Prognose der Strukturdaten des NPVM einzusetzen.

Wechselwirkungen zwischen Raum und Verkehr

Die Wechselwirkungen zwischen Änderungen der Verkehrs- und Raumentwicklung sollen in einer integrierten Modellumgebung abgebildet werden. Umzugsentscheide sind u.a. abhängig von der Erreichbarkeit der Wohn- und Arbeitsplatzstandorte und die Verteilung von Bevölkerung und Arbeitsplätzen wirkt ihrerseits wieder auf die Verkehrserzeugung sowie die Zielwahl im Verkehrsmodell. Die in der Verkehrsmodellierung des UVEK (VM-UVEK) parallel entwickelten und teilweise bereits aufeinander abgestimmten Modelle *Nationales Personenverkehrsmodell* (NPVM) und *Facility Location Choice Simulation* (FaLC) bilden dazu die Grundlage.

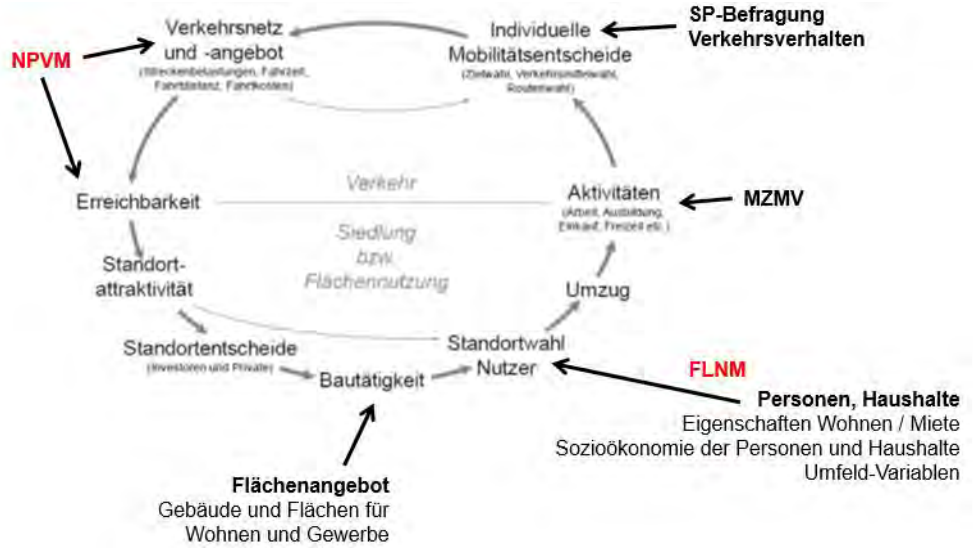


Abbildung 8. Interaktion Raum und Verkehr: Modelle und Datengrundlagen (NPVM: Nationales Personenverkehrsmodell, FLNM: Flächennutzungsmodell, MZMV: Mikrozensus Mobilität und Verkehr, SP-Befragung: Stated Preference Befragung zur Verkehrsmittel- und Routenwahl)

Über die in FaLC generierte und für zukünftige Zeitpunkte prognostizierte SynPop besteht mit Blick auf die Verkehrserzeugung bereits heute eine Schnittstelle zwischen den Modellen. Die SynPop wird zu verhaltenshomogenen Personengruppen zusammengefasst und bietet damit eine Eingangsgrösse für das Verkehrsmodell. Eine vollständige Kopplung der Modelle geht darüber hinaus: Änderungen im Verkehrsangebot induzieren Änderungen im Standortwahlverhalten von Unternehmen und Haushalten; gleichzeitig beeinflusst ein erweitertes oder reduziertes Siedlungs- und Flächenangebot das Standortwahlverhalten und damit die räumliche Verteilung der Nachfrage – die wiederum relevant für die Abbildung der Verkehrsnachfrage ist. Dies ist das mittelfristige Ziel des ARE.

Validierung und Umgang mit SynPop

Die Erwartungen betreffend der Kalibration und Validierung auf der jeweiligen Raumebene können wie folgt zusammengefasst werden:

- Bevölkerung (Altersgruppen) und Arbeitsplätze (Branchen): Adresse, Hektare, Verkehrszone;
- Besitz Mobilitätswerkzeuge (Personenwagen, GA-Besitz, Halbtax-Besitz, Verbundabobesitz): Verkehrszone;
- Personenwagen-Verfügbarkeit: Kanton gekreuzt mit Gemeindetyp, (MS-Regionen);
- Verfügbares Einkommen (Netto): Schweiz, Kanton, (Gemeinden);
- Gekreuzte Eigenschaften (z.B. Haushaltstyp + Anzahl Mobilitätswerkzeuge + Einkommen): Schweiz, Kanton gekreuzt mit Gemeindetyp.

Die Verwendung von SynPop in der Politikberatung bringt nicht nur Herausforderungen betreffend der Qualitätskontrolle sondern auch hinsichtlich der Kommunikation mit sich. Die vermeintliche Genauigkeit einer SynPop kann zu Interpretationen und Anwendungswünschen führen, die über die Möglichkeiten

der SynPop hinausgehen. Folgende Grundsätze müssen aus Sicht des ARE deshalb zwingend berücksichtig werden:

- Die SynPop muss alle offiziellen Datensätze des Bundes einbeziehen und deren Charakteristiken im Ist-Zustand abbilden. Dies ist zentral für die Glaubwürdigkeit der SynPop.
- Die Resultate der SynPop müssen stabil und reproduzierbar sein, es muss EIN Abbild der Schweizer Bevölkerung erstellt werden können. Intervalle und Verteilungen werden für die Validierung und Qualitätskontrolle verwendet.
- Es muss klar kommuniziert werden, bis zu welcher Aggregationsebene die Resultate plausibilisiert sind und weiterverwendet werden können.

Nächste Schritte

Zwischen 2017 und 2019 erstellt das ARE das NPVM neu auf das Basisjahr 2016. Neben einer neuen detaillierteren Zonenstruktur findet u.a. eine stärkere Segmentierung der Verkehrsnachfrage in ausdifferenzierte Personengruppen statt. Dazu muss die SynPop 2016 aus FaLC die notwendigen Variablen in hoher Qualität, Aktualität und räumlicher Auflösung zur Verfügung stellen, so dass über die Aggregation auf die im NPVM vorgesehene Differenzierung der Nachfrage valide Grundlagen für die Verkehrsmodellierung bereitstehen. Weiter erfolgt die Prognose der SynPop auf einen Zustand 2030.

4. Einsatzmöglichkeiten und Herausforderungen

4.1 Einsatzmöglichkeiten von synthetischen Populationen

Teilsynthetische und synthetische SynPop. Eine grundlegende Erkenntnis des Workshops war, dass die verfügbaren Datenmengen über die zu synthetisierende Population die gesamte Vorgehensweise bei der Erstellung der SynPop sehr stark vorgeben. Die vorgestellten SynPop aus dem Ausland mussten teilweise mit räumlich sehr groben oder von der Stichprobengrösse her sehr kleinen Dätensatzen auskommen (siehe Kapitel 2.4 und 2.5). Die Problemstellung bei diesen Projekten war eine grundlegend andere, als bei den Beispielen aus der Schweiz und Österreich (siehe Kapitel 2.1 bis 2.3), wo die zugrundeliegenden Daten sehr umfangreich sind. Weil sich die Herangehensweisen stark unterscheiden, kann es hilfreich sein, die Bezeichnung *synthetische Population* für den ersten und *teil-synthetische Population* für den zweiten Fall zu verwenden. Bei synthetischen Populationen, welche auf kleine Stichproben beruhen, muss man sich der Grenzen der zugrundeliegenden Statistiken bewusst sein. «Zusammenhänge» können meist nur über Expertenwissen und Schätzungen eingebracht werden. Bei teil-synthetischen Populationen sind «Zusammenhänge» teilweise auch bereits in den Ausgangsdatensätzen vorhanden. In diesen Fällen sind Imputationsverfahren und Verknüpfungsmodelle so zu wählen, dass die Zusammenhänge erhalten bleiben.

Anwendungsgebiete von SynPop und Spezialwissen. Der Workshop hat auch gezeigt, dass jede der vorgestellten SynPop über eine Spezialisierung verfügt, und in dieser Spezialisierung besser ist als die anderen SynPop. Dies betrifft einerseits die Verknüpfungsmodelle, welche spezialisiert sind auf das Abbilden von «Zusammenhängen» (konditionale Verteilungen von Variablen), anderseits den Einsatz von zusätzlichen Datensätzen und Expertenwissen, welches über die Grunddatensätze hinausgeht. Es wird spannend sein zu beobachten, wie in den kommenden Jahren wohl einerseits neue SynPop hinzukommen, es anderseits – methodisch gesehen – zu «Fusionen» von SynPop (methodische Vereinheitlichung der grundlegenden Bevölkerungsdatensätze) kommen wird. Möglicherweise bildet sich dabei eine Kern-SynPop heraus, an welcher einzelne Teams weitere Attribute – je nach Fragestellung – anspielen könnten.

Die Diskussionen am Workshop zeigten mehrere wesentliche Vorteile beim Einsatz von SynPop auf:

- **Kleinere Konfidenzintervalle dank imputierten, synthetischen Volldatensätzen:** SynPop können dank Imputation eines Stichproben-Datensatzes genauere Auswertungen erlauben auch dort, wo die Stichprobe selber zu klein wird für statistische Analysen.
- **Nutzung aller vorhandenen Daten:** Im Gegensatz zu Top-Down-Modellen, welche inhärent mit Rundungseffekten infolge der Bildung von Kategorien, Klassen und dergleichen einhergehen, erlauben SynPop als Bottom-Up-Ansatz die Verwendung aller vorhandenen Daten, ohne dass Annahmen zu ihrer statistischen Verteilung nötig sind oder Kategorienbreiten und -grenzen festgelegt werden müssen.
- **Abbildung von Zusammenhängen:** SynPop können – dank der Verknüpfung verschiedener Datensätze und der Abbildung der Zusammenhänge zwischen diesen Datensätzen – für verschiedenste Projekte genutzt werden. Das

detailliertere Wissen über die Bedürfnisse und Charakteristiken der Bevölkerung liefert bessere Entscheidungsgrundlagen auf den Gebieten, welche durch die SynPop abgedeckt werden.

Wie für die verschiedenen SynPop in Kapitel 2 aufgezeigt wurde, werden SynPop heute bereits verwendet im Bereich der Raumentwicklung, der Entstehung und Abwicklung von Mobilität sowie für das Ausgabeverhalten der Bevölkerung. Auch, wieviel Energie (z.B. Wärme) in einem gewissen Stadtteil verwendet wird, lässt sich – wenn man zur Kalibrierung auch über Absatzdaten z.B. für die gesamte Stadt verfügt – über die Summierung des abgeschätzten Verbrauchs aller Einzelgebäude abschätzen.

Neue Einsatzmöglichkeiten von SynPop können sein:

- Räumliche oder soziale Verteilungseffekte von Politikinstrumenten, z.B. von Lenkungsabgaben auf Energie und deren Rückverteilung an Arbeitgeber und die Bevölkerung.
- Künftige Weiterentwicklung der Bevölkerung hinsichtlich Einkommen und Einkommensverwendung, für die Abschätzung der Steuerkraft einzelner Gemeinden, Regionen und Kantone, für Fragen für die Weiterentwicklung des Finanzausgleichs und der Lastenverteilung zwischen verschiedenen staatlichen Ebenen.
- Auch bei Fragestellungen, welche nur einen schwach ausgeprägten Raumbezug haben, können künftig SynPop eingesetzt werden, weil sie eine Mikrosimulation unter Verwendung aller vorhandenen Ausgangsdaten ermöglichen. Dies betrifft beispielsweise Fragen im Bereich der nationalen Sozialversicherungen sowie der Finanzierung des Gesundheitswesens, wo sich Bund und Kantone die Aufgaben teilen.

4.2 Qualitätskontrolle von synthetischen Populationen

Die am Workshop vorgestellten SynPop sind alle zuerst als forschungsnahes Projekt entstanden. Qualitätsprüfungen und Qualitätskriterien wurden nicht vorgängig festgelegt. Ob die SynPop die für den jeweiligen Anwendungsbereich notwendigen «Zusammenhänge» wirklich abbildet, wird meist nicht systematisch überprüft. Interessant wird es sein, wie sich hier in Zukunft Verfahren und Projektschritte zur Definition, Überprüfung und Abnahme der geforderten Qualität von SynPop – jeweils spezifisch für einen gegebenen Anwendungsbereich – etablieren werden. Dabei können fünf Stufen unterschieden werden:

1. **Qualitätsprüfung im engeren Sinn**: Überprüfung der einzelnen Datensäte und der vorkommenden Zahlenwerte. Dabei ist insbesondere darauf zu achten, dass verschiedene Datensätze unterschiedliche Codierungen für fehlende Werte verwenden (durch leer lassen des Datenfeldes, oder mit numerischen Werten wie «–1», «–999», usw.), einzelne Daten aus Gründen des Datenschutzes maskiert sein können, oder es sich um kategoriale Daten handelt.
2. **Vergleich der Dichteverteilungen mit den ursprünglichen Datensätzen**: Wo nicht Echtdatensätze verwendet werden, sondern Stichprobendatensätze imputiert wurden, kann die Dichteverteilung der SynPop mit jener der zugrundeliegenden Stichprobe verglichen werden. Hiermit können jene Hochrechnungs- und Imputationsverfahren identifiziert werden, welche nicht verteilungstreu sind.

3. **Vergleich mit aggregierten Grössen aus anderen Datenquellen:** Die Gesamtzahl der Bevölkerung, der Gebäude, des Konsums, der Verkehrsleistung, usw., kann verglichen werden mit Angaben aus anderen Datenquellen. Möglicherweise gehen die anderen Datenquellen zwar auf das gleiche Datenregister zurück, verwenden aber andere Systemgrenzen. Dies kann bereits unterschiedliche Definitionen aufdecken, z.B. welche Kategorien von ausländischen Staatsangehörigen zur ständigen Wohnbevölkerung gezählt werden. Möglicherweise werden auch die Daten von Privatpersonen und juristische Personen je nach Datensatz anders behandelt. Durch juristische Personen gehaltene Motorfahrzeuge und ihre Verkehrsleistung können beispielsweise in bestimmten Datensätzen eingeschlossen sein, aus anderen aber ausgeschlossen.
4. **Validierung mit einem Teilsample:** Bei der Bildung der SynPop einen Teil der verwendeten Echtdaten (z.B. alle Daten für eine mittelgrosse Stadt) weglassen, und danach die SynPop-Ergebnisse für diese Stadt mit den Echtdaten vergleichen.
5. **Formulierung und Prüfung von experten-basierten Hypothesen:** Dieses Verfahren eignet sich besonders, wenn eine SynPop Sachverhalte abbildet, zu welchen es kaum bis gar keine realen Beobachtungen gibt (was in den meisten Fällen das Ziel einer SynPop ist). Dazu formulieren Experten zuerst Hypothesen («Bevölkerungssegment mit höheren Einkommen sollten eine überdurchschnittliche Jahresmobilität aufweisen; im ländlichen Raum jedoch weniger ausgeprägt als im städtischen»). Diese können durch Auswertung der SynPop bestätigt oder verworfen werden. Wichtig ist, dass Entwickler und Anwender einer SynPop die Hypothesen vor der Entwicklung der SynPop gemeinsam formulieren und festhalten.

4.3 Herausforderungen und Ausblick

Neben den methodischen Aspekten (voll- oder teilsynthetische Ansätze, je nach der Datenverfügbarkeit) und der Qualitätskontrolle, welche in Kap. 4.2 behandelt wurden, stehen für die Weiterentwicklung der SynPop die folgenden Handlungsfelder im Fokus.

Anwendungsbereiche einer SynPop. Die erstmalige Erstellung einer SynPop bedingt einen hohen Initialaufwand. Deshalb wird meistens eine SynPop anfänglich sehr stark auf einen bestimmten Anwendungszweck hin entwickelt, das heisst, dass auf einzelne «Zusammenhänge» zwischen Variablen (die konditionalen Verteilungen einer Variable unter Berücksichtigung einer anderen Variable) fokussiert wurde. Kommen dann später zusätzliche Anwendungszwecke hinzu, besteht die Gefahr, dass aus Ressourcengründen nicht alle neu entstehenden Zusammenhänge geprüft und plausibilisiert werden. Deshalb sollte immer deklariert werden, welche Variablen konditionalen Verteilungen genügen und welche nicht, und welche Datensätze *nicht* in die SynPop einfliessen.

Datennutzungsverträge und Projektdauer. Eine SynPop ist immer mit einem beträchtlichen organisatorischen und administrativen Aufwand zur Einhaltung des Datenschutzes (inkl. Vorgaben der Datennutzungsverträge) verbunden. Dieser Aspekt gewinnt nochmals an Bedeutung, falls die Bevölkerungs-Registerdaten nicht aggregiert pro Hektare, sondern adressscharf verwendet werden. Die Datennutzungsverträge gelten deshalb in aller Regel jeweils nur für ein Projekt. Nach Abschluss eines Projektes muss die Löschung der erhaltenen Daten

bestätigt werden. Zurück bleiben die Verknüpfungsmodelle. Die Qualität der Verknüpfungsmodelle kann aber nur zusammen mit den Ausgangsdaten effizient überprüft werden. Wichtig ist deshalb, dass die Datennutzungsverträge nicht nur die Übergabe an den Auftraggeber und dessen Abnahmeverfahren abdecken, sondern auch eine nachfolgende Gewährleistungsfrist. Wurden nach Projektende die Grunddaten gelöscht, ist es für die Abklärung neuer Anwendungsfälle nicht möglich, indikative Auswertungen der SynPop durchzuführen und vorgängig abzuklären, welchen Zusatznutzen die SynPop für einen bestimmten Anwendungsbereich haben wird. Dies hemmt grundsätzlich die methodische Weiterentwicklung von SynPop und ihrer Verknüpfungsmodelle. Wichtig wäre deshalb, nach Abschluss eines Projekts zwar die Ausgangs-Grunddatensätze sofort löschen zu müssen, Teile der erzeugten SynPop aber erhalten zu dürfen (auch wenn darin Teile der Grunddatensätze enthalten sind), um Analysen für mögliche Anwendungszwecke in zu definierendem Umfang vornehmen zu dürfen.

Ständige Aktualisierung einer SynPop. Die Grunddatensätze liegen immer wieder in neuen Versionen vor. In der Schweiz wäre es grundsätzlich sinnvoll, eine SynPop in Jahresschritten neu zu bilden, weil die Echtdatensätze ständig (tagesaktuell) weitergeschrieben werden und auch in jedem Jahr neue Datenstände bei einzelnen Stichprobendatensätzen vorliegen (z.B. Strukturerhebung jährlich; HABE alle 3 Jahre, MZMV alle 5 Jahre). Wichtig ist deshalb auf Seiten der SynPop-Entwickler eine strikte Trennung zwischen Ausgangsdaten und Verknüpfungsmodellen, so dass bei Vorliegen einer neuen Datenversion die SynPop «auf Knopfdruck» neu generiert werden kann. Auf Seiten der Datenanbieter (in der Schweiz in den meisten Fällen das BFS) ist es wichtig, dass sich die Struktur und die Datenformate von Grunddatensätzen nicht ohne zwingenden Grund ändern.

Prognosefähige SynPop. Für eine aussagekräftige Prognosefähigkeit benötigt eine SynPop zahlreiche Annahmen zur zeitlichen weiteren Entwicklung der in den Registerdaten und Stichprobendatensätzen abgebildeten Verhältnisse. Wie werden die Bevölkerung, die Einkommen, das Mobilitätsverhalten, der Gebäudebestand und die Betriebe fortgeschrieben, mit welcher räumlichen Auflösung und welchen gegenseitigen Abhängigkeiten? Wie bildet man dabei die zunehmende Unsicherheit ab? Es gibt zwei mögliche Ansätze für den Blick in die Zukunft: Entweder man betrachtet jede Zeitscheibe separat (die verschiedenen Ausgangsdatensätze für die SynPop werden einzeln für das gewünschte Prognosejahr hochgerechnet oder geschätzt, und daraus wird dann – methodisch gesehen genau gleich wie bei einem Bezugsjahr in der Vergangenheit – die SynPop gebildet), oder man kreiert eine Fortschreibung über die Jahre basierend auf Evolutionsmodellen (dazu braucht es die SynPop für ein Ausgangsjahr, und dann Modelle, welche Geburten, Todesfälle sowie Änderungen der Haushalte, Wohn- und Arbeitsorte in Jahresschritten beschreiben). Im Gegensatz zu den Grunddatensätzen gibt es zu keinem der beiden Verfahren eine «common practice» oder allgemein verfügbare Annahmen. Wichtig wäre daher, öffentlich finanzierte Entwicklungen solcher Annahmen, namentlich mit räumlicher Auflösung, öffentlich zugänglich zu machen. Das ARE z.B. trägt aktiv dazu bei. Unter Einhaltung von Daten- und Eigentumsschutz dokumentiert und publiziert das ARE Annahmen, Methoden und Resultate von Prognosen transparent.

Für Anwendungen in der Verwaltung ist wichtig, dass die Perspektivarbeiten des Bundes (Bevölkerungsentwicklung des BFS, Wirtschaftsentwicklung des seco) als Rahmenentwicklung verwendet werden.

Bildung einer Community. Die Entwicklung von synthetischen Populationen wird immer mit einem Aufwand verbunden sein, welcher die Beteiligten zur Effizienz und zur Zusammenarbeit zwingt. Da jeder Anwendungszweck und jeder Grunddatensatz Spezialwissen erfordert, werde sich dabei Spezialisierungen herausbilden.

Die Auswertung des Fragebogens hat gezeigt, dass der Workshop auf grosses Interesse gestossen ist und weiterer Bedarf nach Austausch besteht. Das ARE wird die methodischen Entwicklungen weiterverfolgen und auch weitere Erfahrungen in der Anwendung von SynPop sammeln. Sobald neue Erkenntnisse und/oder Erfahrungen in der Community vorliegen wird das ARE einen zweiten Workshop - voraussichtlich Ende 2019 - zum Thema der SynPop organisieren.

5. Liste der Workshop-Teilnehmenden

Name:	Vorname:	Institution:
Arendt	Michael	Arendt Consulting
Auf der Maur	Alex	Prognos AG
Bernet	Aurelius	BERNET-Engineering
Brändle	Thomas	Eidgenössische Finanzverwaltung EFV
Brunner-Patthey	Olivier	Office fédéral des assurances sociales OFAS
Cataldi	Damien	Direction générale des transports, Canton de Genève
Catillaz	Andreas	Bundesamt für Umwelt BAFU
Cavallasca	Lorenzo	Tiefbauamt, Stadt Zürich
Cerri	Valérie	Service des transports, Canton de Neuchâtel
Colombier	Carsten	Eidgenössische Finanzverwaltung EFV
Cotter	Stéphane	Office fédéral de la statistique OFS
Danalet	Antonin	Bundesamt für Raumentwicklung ARE
Erath	Alexander	Erath Rusterholtz, van Eggermond & Co
Erni	Kurt	Amt für Verkehr und Tiefbau, Kanton Solothurn
Faust	Anne-Kathrin	Bundesamt für Energie BFE
Finné	Gordon	Departement Bau, Verkehr und Umwelt, Kanton Aargau
Frömelt	Andreas	ETH Zürich
Heemann	Detlef	Roland Müller Küsnacht AG
Köglmaier	Andreas	PTV Group
Kost	Michael	Bundesamt für Energie BFE
Lieberherr	Johannes	ttools gmbh
Madl	Edith	Bundeskanzlei BK
Marini	Marcello	ETH Zürich
Marti	Res	Fachstelle für Statistik, Kanton Zug
Martínez	Adrian	Eidgenössische Finanzverwaltung EFV
Métrailler	Denis	Schweizerische Bundesbahnen SBB AG
Müller	André	Ecoplan AG
Müller	Kirill	datatools GmbH
Nökel	Klaus	PTV Group
Ordon	Christian	Amt für Verkehr, Kanton Zürich
Orhan	Özkul	Amt für Verkehr, Kanton Zürich
Peters	Rudi	Administration fédéral des contributions AFC
Roulin Perriard	Anne	Bundeskanzlei BK
Salamin	Paul-André	Office fédéral des assurances sociales OFAS
Scherr	Wolfgang	Schweizerische Bundesbahnen SBB AG
Schiller	Christian	TU Dresden
Schwyn	Markus	Bundesamt für Statistik BFS
Stetter	Adrian	EBP Schweiz AG
Tinguely	Martin	Bundesamt für Strassen ASTRA
Tschopp	Martin	Bundesamt für Raumentwicklung ARE
Villiger	Simon	Fachstelle für Statistik, Kanton Zug
Vitins	Basil	ASE (Analysis Simulation Engineering) AG
Vrtic	Milenko	TransOptima GmbH
Weis	Claude	TransOptima GmbH

Referenten

Balmer	Michael	Senozon AG
Bierlaire	Michel	EPFL, Lausanne
Bodenmann	Balz	Strittmatter Partner AG
Cyganski	Rita	DLR Institut für Verkehrsforschung
de Haan	Peter	EBP Schweiz AG
Haupt	Thomas	Senozon AG
Justen	Andreas	Bundesamt für Raumentwicklung ARE
Mathys	Nicole	Bundesamt für Raumentwicklung ARE
Moser	Peter	Statistisches Amt, Kanton Zürich
Müller	Michael	EBP Schweiz AG

6. Literaturverzeichnis

Beckman R.J., Baggerly K.A., McKay M.D. (1996): Creating synthetic baseline populations. Transportation Research Part A 30 (6), 415–429.

BBSR (2012): Raumordnungsprognose 2030. Bevölkerung, private Haushalte, Erwerbspersonen. Analysen Bau.Stadt.Raum Bd.9, Bonn.

BMVI (2018): Bundesministerium für Verkehr und digitale Infrastruktur: Mobilität in Deutschland (MiD), bundesweite Befragung von Haushalten zu ihrem alltäglichen Verkehrsverhalten, http://www.mobilitaet-in-deutschland.de/

Bundesamt für Raumentwicklung ARE (2014): Entwicklung eines Flächennutzungsmodells für die Schweiz http://www.are.admin.ch/flnm

Bundesamt für Raumentwicklung ARE (2017): Weiterentwicklung Flächennutzungsmodellierung: Wohnstandortwahl: Erweiterung des Modells FaLC: Verhaltensmodelle und synthetische Population; http://www.are.admin.ch/flnm

Farooq B., Bierlaire M., Hurubia R., Flötteröd G. (2013): Simulation based population synthesis. Transportation Research Part B, Volume 58, December 2013, Pages 243-263.

Heinrichs, M., Krajzewicz, D., Cyganski, R. & von Schmidt, A. (2016): Introduction of car sharing into existing car fleets in microscopic travel demand modelling, in: Personal and Ubiquitous Computing, S. 1-11, Springer. DOI: https://doi.org/10.1007/s00779-017-1031-3, ISSN 1617-4909, 2017.

Heinrichs, M., Krajzewicz, D., Cyganski, R. & von Schmidt, A. (2016): Disaggregated Car Fleets in Microscopic Travel Demand Modelling, In Procedia Computer Science, Volume 83, S. 155-162, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2016.04.111.

Heldt, B., Donoso, P., Bahamonde-Birke, F., & Heinrichs, D. (2017): Estimating bid-auction models of residential location using census data with imputed household income. Accepted for Journal of Transport and Land Use.

Horni, A., K. Nagel and K.W. Axhausen (eds.) (2016): The Multi-Agent Transport Simulation MATSim, Ubiquity, London. DOI: https://doi.org/10.5334/baw

Martínez, F., & Donoso, P. (2010): The MUSSA II land use auction equilibrium model. In F. Pagliara, J. Preston, & D. Simmonds (Eds.), Residential Location Choice (S. 99-113): Springer.

Moekel, R. (2016): Constraints in household relocation: Modeling land-use/transport interactions that respect time and monetary budgets. Journal of Transport and Land Use, Vol. 10(2), S. 1-18.

Müller, K., Axhausen, K. W. (2011): Population Synthesis for Microsimulation: State of the Art. Papers presented at the 90th Annual Meeting of the Transportation Research Board, Washington, D.C, January 2011.

Prichard, D. & Miller, E. J. (2012): Advances in population synthesis: fitting many attributes per agent and fitting to household and person margins simultaneously. Transportation, 39 (3), S. 685-704.

Statistisches Bundesamt 2018. Mikrozensus Verkehr. https://www.destatis.de/DE/ZahlenFakten/GesellschaftStaat/Bevoelkerung/Mikrozensus.html

TU Dresden (2018): Forschungsprojekt Mobilität in Städten – SrV, https://tu-dresden.de/die_tu_dresden/fakultaeten/vkw/ivs/srv

von Schmidt, A., Cyganski, R. & Krajzewicz, D. (2017): Generierung synthetischer Bevölkerungen für Verkehrsnachfragemodelle - Ein Methodenvergleich am

Beispiel von Berlin, in: HEUREKA'17 - Optimierung in Verkehr und Transport, S. 193-210, FGSV-Verlag, ISBN 978-3-86446-177-4, 2017.

Ye, X., Konduri, K., Pendyala, R. M., Sana, B., Waddel, P. (2009): A methodology to match distributions of both household and person attributes in the generation of synthetic populations. Paper presented at the 88th Annual Meeting of the Transportation Research Board, Washington, D.C., January 2009.

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Eidgenössisches Departement für
Umwelt, Verkehr, Energie und Kommunikation UVEK

Bundesamt für Raumentwicklung ARE
Sektion Grundlagen

SynPop

Synthetische Populationen für die Politikberatung in der Schweiz

Bern, 08.12.2017
Nicole Mathys, Andreas Justen, ARE Sektion Grundlagen

Motivation der heutigen Tagung

- Anforderungen der Politik an Planungsgrundlagen sind hoch
- Datensätze:
  - Letzte Volkszählung im 2000
  - Neues Statistiksystem: Register (Vollerhebungen) werden durch Strichprobenerhebungen ergänzt
- Agentenbasierte Modellierungen mit vielversprechenden Möglichkeiten
- ARE hat erste Erfahrungen gesammelt

Ziele der heutigen Tagung

- Gemeinsames Verständnis
  - Was verstehen wir unter einer einer SynPop?
  - Für welche Fragestellungen sind SynPop hilfreich?
  - Welche Unterschiede bestehen zwischen den bestehenden Synpop?
- Best practice
  - Austausch zu Erfahrungen: Möglichkeiten und Bedürfnisse
  - Umgang mit spezifischen Bedürfnissen der Politikberatung
  - Empfehlungen für Weiterentwicklungen
- Netzwerk:
  - Übersicht über Akteure: Ersteller und Nutzer
  - Anwendungsbereiche aufzeigen
  - Synergien nutzen

Programm Vormittag

09.15 Uhr	Begrüssung, Ziele der Tagung	Dr. Nicole Mathys
09.30 Uhr	Simulation-based population synthesis using Gibbs sampling – the Brussels case	Prof. Dr. Michel Bierlaire
10.00 Uhr	Aufbau und Anwendung einer synthetischen Bevölkerung im Verkehrsmodell Oberösterreich	Thomas Haupt
10.30 Uhr	*Kaffeepause*	
10.50 Uhr	Synthetische Populationen aus FaLC-sim für das Nationale Personenverkehrsmodell	Dr. Balz Bodenmann
11.20 Uhr	Wie schweizerisch ist die synthetische Schweiz von EBP?	Dr. Michel Müller
12.00 Uhr	*Stehlunch bis 13.00 Uhr*	

Programm Nachmittag

13.00 Uhr	Wrap–up Vormittag	Dr. Peter de Haan
13.15 Uhr	Generierung synthetischer Bevölkerungen für Berlin – Möglichkeiten und Grenzen	Rita Cyganski
13.45 Uhr	Qualitätssicherung bei synthetischen Bevölkerungen	Dr. Peter Moser
14.15 Uhr	Bedürfnisse der Bundesverwaltung & Einsatz in den Themen Raum und Verkehr	Dr. Andreas Justen
14.45 Uhr	Abschlussdiskussion & Ausblick	Dr. Peter de Haan Dr. Nicole Mathys
15.30 Uhr	*Ende der Tagung*	

Simulation-based population synthesis using Gibbs sampling

Bilal Farooq[1] Michel Bierlaire[2]

[1]Civil Engineering
Ryerson University

[2]Transport and Mobility Laboratory
School of Architecture, Civil and Environmental Engineering
Ecole Polytechnique Fédérale de Lausanne

December 8, 2017

Outline

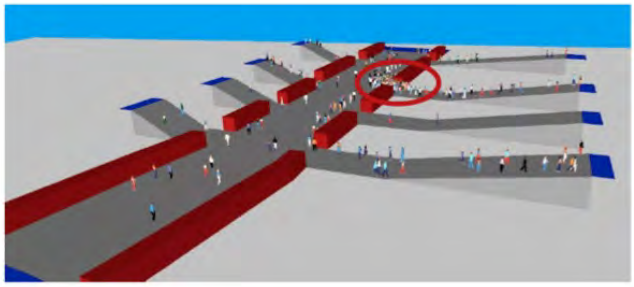
Modelling and Micosimulation

Urban area

$$\Pi = \sum_{i=1}^{N} \frac{\gamma_i}{\alpha_i} \left\{ \left(f^r(X_i^r) - f^c(X_i^c) \right)^{\vartheta} \left(\left(\frac{q_i}{\gamma_i} + 1 \right)^{\alpha_i} - 1 \right) \right\} + f^z(z)$$

Mobility Hub

$$Q(\xi,\tau) = \begin{cases} n(\xi,\tau)\left\{1 - \exp\left[-\gamma_\xi A_\xi \left(\frac{1}{n(\xi,\tau)} - \frac{1}{N_\xi}\right)\right]\right\} & \text{if } 0 < n(\xi,\tau) < N_\xi \\ 0 & \text{otherwise} \end{cases}$$



TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Agent based Microsimulation

Policy scenarios

Agent based Microsimulation

$\text{Time}_{\text{start}}$ ..... Time_{end}

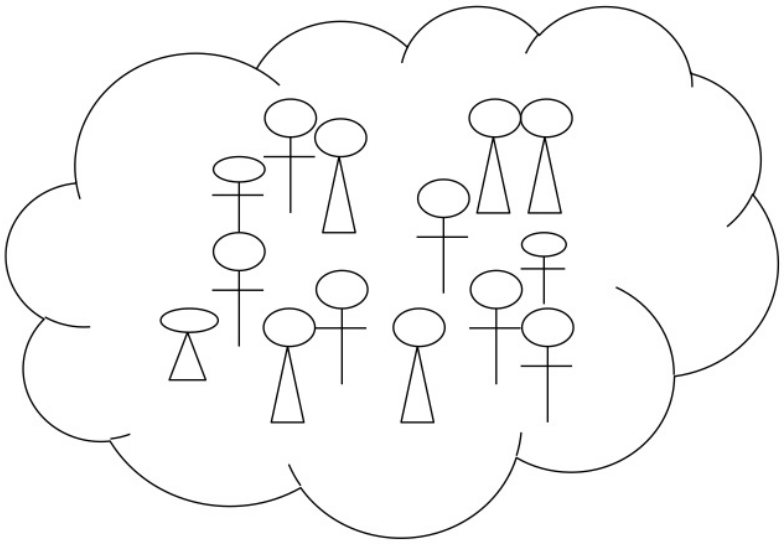
Future Indicators

- Accessibility
- Prices
- Energy usage
- Emissions
- Population densities
- Demographics
- Economy

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Population Synthesis



SustainCity project

- European Union funded mega research project
- More than 10 major European universities involved
- Aims:
  - Integrated land use and transportation modelling framework
  - Demographics, environment, and multi-scale issues
- Case studies
  - Paris
  - Zurich
  - Brussels

SustainCity: Brussels case study [Farooq et al., 2015]

Brussels case study

- Data sources (extremely limited)
  - Incomplete conditionals of households and persons (Census 2001)
  - Travel survey of households and individuals (MOBEL 1999)
    - **3063 observations (0.2%)**
- Synthetic household attributes
  - Size, children, workers, cars, income, university education, dwelling type, sector

Brussels case study

- Data sources (extremely limited)
  - Incomplete conditionals of households and persons (Census 2001)
  - Travel survey of households and individuals (MOBEL 1999)
    - **3063 observations (0.2%)**
- Synthetic household attributes
  - Size, children, workers, cars, income, university education, dwelling type, sector

- *Conventional synthesis procedures were not usable*

Evolution of Synthesis Methods in Transport

Initial efforts

- From *Four-Stage* to *Activity based Integrated* modelling
- Forecasting behaviour using individual level models
- Synthesis for TRansportation ANalysis SIMulation System (TRANSIMS) [Beckman et al., 1996]

Evolution of Synthesis Methods in Transport

Initial efforts

- From *Four-Stage* to *Activity based Integrated* modelling
- Forecasting behaviour using individual level models
- Synthesis for TRansportation ANalysis SIMulation System (TRANSIMS) [Beckman et al., 1996]

Existing approach

- **Fitting based approach**
  - Iterative proportional fitting
    - By far the most commonly used approach
  - Combinatorial optimization
- Adjusting sample weights to fit the aggregate statistics

Iterative Proportional Fitting (IPF) [Beckman et al., 1996]

- Contingency Table (CT) from sample
  - Categorization of variables of interest
  - Totals for each cell of the resulting multi-way table
- Fitting: Multi-constraint gravity model sort of formulation
  - Sample used to initialize the contingency table
  - Use marginal as dimensional totals
  - Adjust the cell proportions to fit dimension totals
  - Iterate while the error is large
  - Odd-ratio is maintained
- Generation of agents based on fitted weights
  - Monte Carlo simulation for fractions

Combinatorial Optimization (CO) [Williamson et al., 1998]

- Zone-by-zone
- 0-1 weights for each row in the sample
- Optimizing the weights to fit zonal marginals
- Use of hill-climbing, simulated annealing, and genetic algorithm to estimate the best set of obs. weights for each zone

Key issues

- Optimization resulting in one synthetic population
  - Data are incomplete and purposely tampered with sophisticated anonymizing techniques
  - There can be any number of solutions
- Cloning of data rather than creation of a heterogeneous representative population
- Focus on fitting marginals
  - Generation of correct correlation structure is more important, as that is what the behavioural models are operating on

Key issues

- Over reliance on the accuracy of the microdata, without serious consideration to the sampling process and assumptions
- Large enough sample size
- Inefficient use of the available data
- Discrete agent attributes only
- Scalability issues

Problem statement

- True population: Individual agents defined as a set of attributes $X = (X^1, X^2, ..., X^n)$
  - Discrete (e.g. marital status) or continuous (e.g. income)
  - Unique joint distribution represented by $\pi_X(x)$
- No direct access to $\pi_X(x)$ and hard to draw from
- Instead, only partial views of $\pi_X(x)$
  - Marginals, conditional-marginals, and samples

Problem statement

- Develop a synthesis procedure that lets us use these views to draw a synthetic population as if we were drawing from $\pi_X(x)$
  - At the same time, ensuring that the empirical distribution $\pi_{\hat{X}}(\hat{x})$ of $\hat{X}$ resulting from the realized synthetic population is as close to $\pi_X(x)$ as possible

Simulation based approach [Farooq et al., 2013]

- Propose to use **Gibbs sampler** for drawing synthetic population
- MCMC method that uses $\pi(X^i|X^j = x^j$, for $j = 1...n$ & $i \neq j) = \pi(X^i|X^{-i})$ for $i = 1, ..., n$ to simulate drawing from $\pi_X(x)$ [Geman and Geman, 1984]
- Key challenge: Preparation of the conditional distributions for attributes from available data sources

Incomplete conditionals

- Full-conditionals rarely available

Completing conditionals by assumptions

- If in $\pi(X^1|X^{-1}) = \pi(X^1|X^{(2\ldots k)}, X^{((k+1)\ldots n)})$ only $\pi(X^1|X^{(2\ldots k)})$ is available
  - In case of no other information, $\pi(X^1|X^{-1}) = \pi(X^1|X^{(2\ldots k)}), \forall X^{((k+1)\ldots n)}$
  - Worst case, we can use $\pi(X^1|X^{-1}) = \pi(X^1)$

Completing conditionals by assumptions

- If in $\pi(X^1|X^{-1}) = \pi(X^1|X^{(2\ldots k)}, X^{((k+1)\ldots n)})$ only $\pi(X^1|X^{(2\ldots k)})$ is available
  - In case of no other information, $\pi(X^1|X^{-1}) = \pi(X^1|X^{(2\ldots k)}), \forall X^{((k+1)\ldots n)}$
  - Worst case, we can use $\pi(X^1|X^{-1}) = \pi(X^1)$

- For $(Age|Sex, Income)$
  - From data only $(Age|Income)$ available
  - Assume that for all values of Sex, $(Age|Sex, Income) = (Age|Income)$
  - No matter the Sex of a person is, Age is only dependent on $Income$

Completing conditionals by domain knowledge

- In case of domain knowledge
  $\pi(X^1|X^{(2\ldots k)}, X^{((k+1)\ldots n)} = a) = \pi^a(X^1|X^{(2\ldots k)})$,
  $\pi(X^1|X^{(2\ldots k)}, X^{((k+1)\ldots n)} = b) = \pi^b(X^1|X^{(2\ldots k)})$,
  ...

Completing conditionals by domain knowledge

- In case of domain knowledge
  $\pi(X^1|X^{(2...k)}, X^{((k+1)...n)} = a) = \pi^a(X^1|X^{(2...k)})$,
  $\pi(X^1|X^{(2...k)}, X^{((k+1)...n)} = b) = \pi^b(X^1|X^{(2...k)})$,
  ...

- For $(Income|Sex, Age)$
  - From data only $(Income|Sex)$ available
  - Known: Infants do not have income, students have low income
    - $(Income|Sex, Age) = \alpha(Income|Sex)$ for $Age = 1...12$
    - $(Income|Sex, Age) = \beta(Income|Sex)$ for $Age = 13...18$
    - $(Income|Sex, Age) = \gamma(Income|Sex)$ for $Age > 18$
    - $\alpha + \beta + \gamma = 1$ and $\alpha < \beta < \gamma$

Completing conditionals by parametric models

- For instance, Logit model $\pi(X_l^1|X_m^{-1}) = \frac{e^{(V_{X_l^1}|X_m^{-1})}}{\sum_{p=1}^{L}\left(e^{(V_{X_p^1}|X_m^{-1})}\right)}$

Completing conditionals by parametric models

- For instance, Logit model $\pi(X_l^1|X_m^{-1}) = \frac{e^{(V_{X_l^1}|X_m^{-1})}}{\sum_{p=1}^{L}\left(e^{(V_{X_p^1}|X_m^{-1})}\right)}$

- For $(Dwelling|Income, Sex, Age)$
  - In sample $(Dwelling, Age, Sex)_p$ for a person are available
  - In zone (z) where person is living
    - Average income by dwelling type (av_inc)
    - ...
  - Dwelling choice model can be estimated for person:
    $dwel\_typ = (attached, semidetached, detached, apartment)$
    and $V^i_{(p,z)} = ASC^i + \beta^i_{age_p} \times Age + \beta^i_{av\_inc_z} \times av\_inc_z + interactions + ...$

Population from Swiss Census

- Access to Swiss Census for 2000
  - Person and household attributes (Except for Income)
- Selected area: postal code in Lausanne
  - CH-1004
  - 28,533 persons
- Four Person attributes (384 combinations)
  - Age (<15, 15-24, 25-34, 35-44, 45-54, 55-64, 65-74, >74)
  - Sex (Female, Male)
  - Household size (1, 2, 3, 4, 5, 6 or more)
  - Education level (none, primary, secondary, university/college)

Comparison between IPF and Simulation

- Criteria: how well the joint distribution is reproduced?

Data preparation

- Prepared same type of datasets as commonly available
  - Individual level microsample
    - Drawing from Census: Uniformly, without replacement
    - No sampling-zero
  - Zonal level conditionals (with various level of completion)
    - By counting from Census

List of available sample sizes

No.	Sample Size
1	20%
2	10%
3	5%
4	3%
5	1%

List of available sample sizes

No.	Sample Size
1	20%
2	10%
3	5%
4	3%
5	1%

- In practice the sample size is 5% or less
- Larger sizes used to investigate representativeness

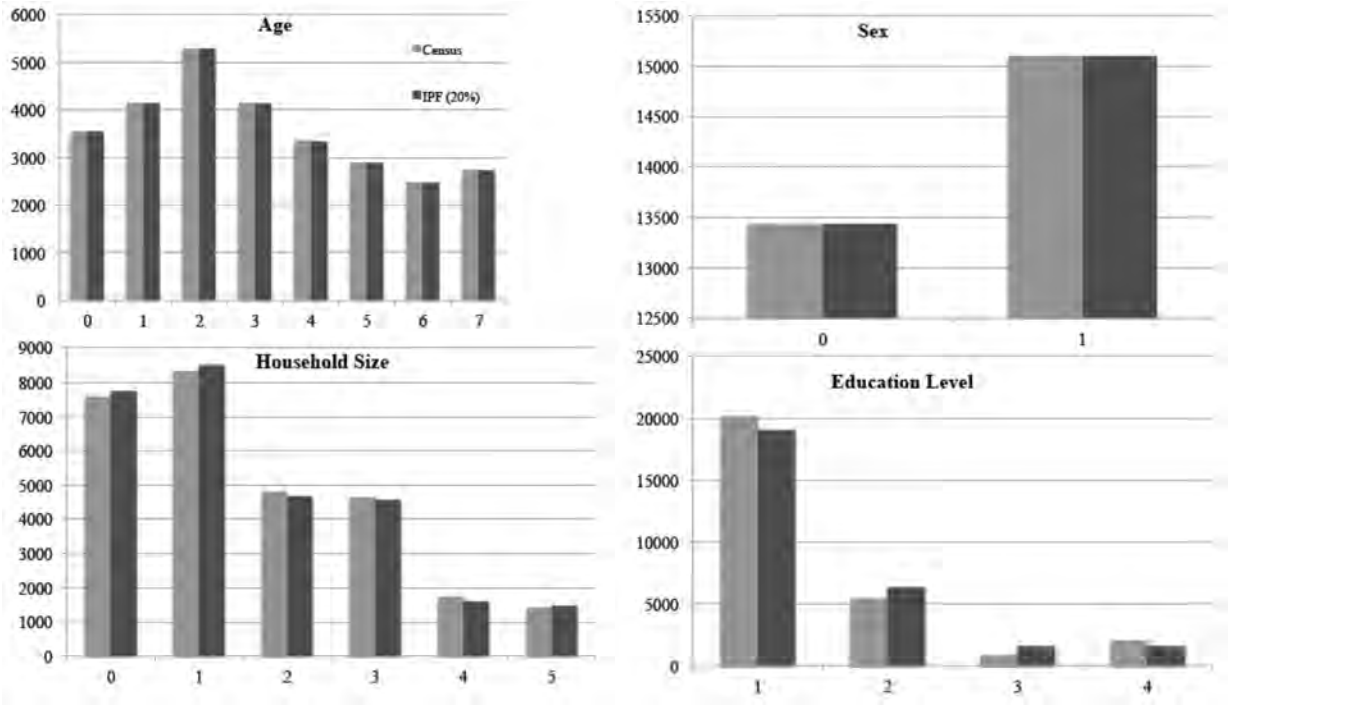
List of available conditionals

No.	ID	Conditionals
1	*FullCond*	$\pi(age\|sex, hhld\_size, edu\_level)$ $\pi(sex\|age, hhld\_size, edu\_level)$ $\pi(hhld\_size\|age, sex, edu\_level)$ $\pi(edu\_level\|age, sex, hhld\_size)$
2	*Partial_1*	$\pi(age\|\cancel{sex}, hhld\_size, edu\_level)$ $\pi(sex\|age, hhld\_size, edu\_level)$ $\pi(hhld\_size\|age, sex, edu\_level)$ $\pi(edu\_level\|age, sex, hhld\_size)$
3	*Partial_2*	$\pi(age\|\cancel{sex}, hhld\_size, edu\_level)$ $\pi(sex\|age, hhld\_size, edu\_level)$ $\pi(hhld\_size\|age, \cancel{sex}, edu\_level)$ $\pi(edu\_level\|age, sex, hhld\_size)$
4	*Partial_3*	$\pi(age\|\cancel{sex}, hhld\_size, edu\_level)$ $\pi(sex\|age, hhld\_size, edu\_level)$ $\pi(hhld\_size\|age, \cancel{sex}, edu\_level)$ $\pi(edu\_level\|age, \cancel{sex}, hhld\_size)$

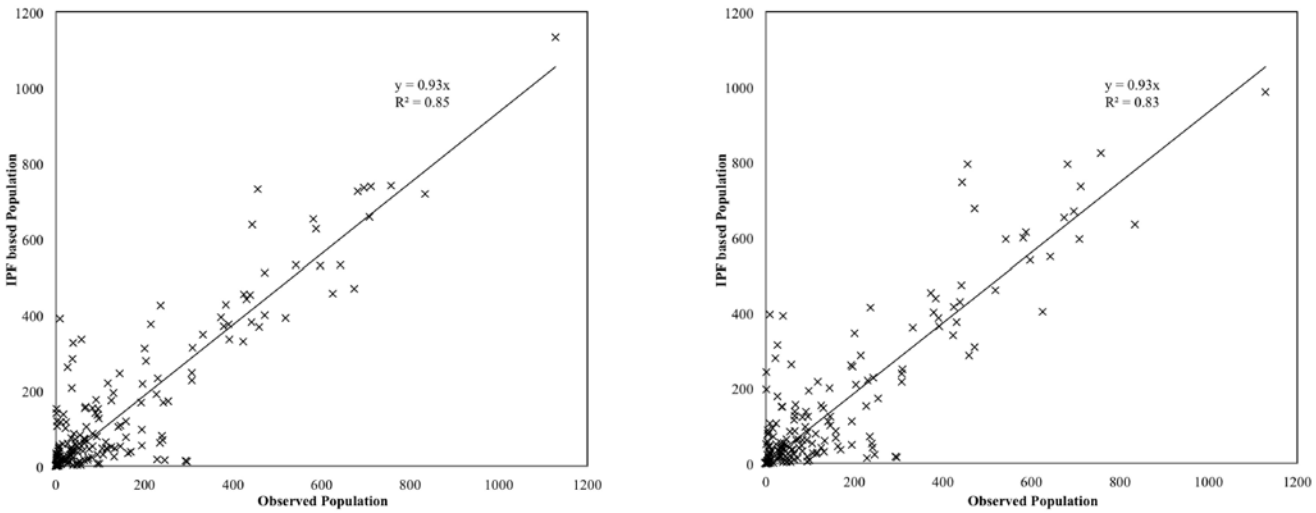
Data preparation

- Based on sample-conditional combinations
  - 20 possibilities
- IPF can use marginals only
  - Number of experiments collapses to 5
- Simulation based synthesis
  - Used conditionals only (used lesser information)
  - Number of experiments collapses to 4

Results: IPF and Census marginals



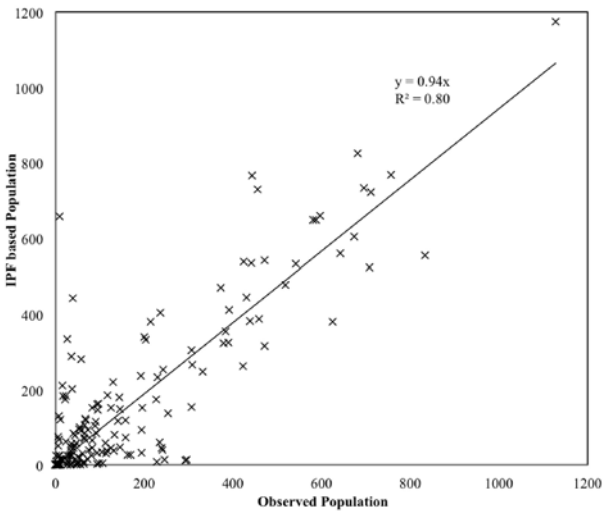
Results: Fit of IPF with Census joint distribution



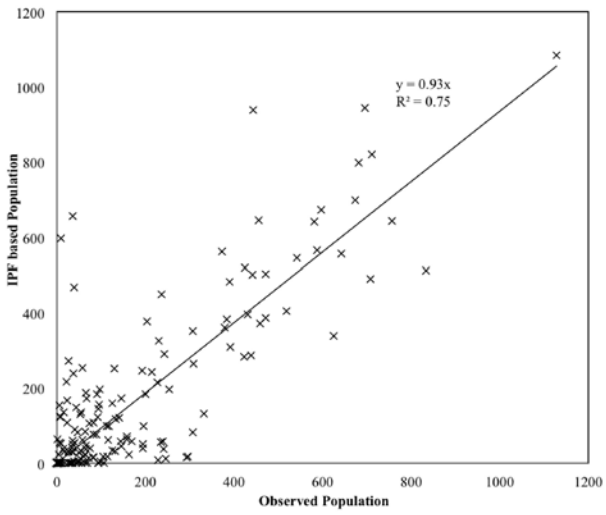
IPF with 20% sample

IPF with 10% sample

Results: Fit of IPF with Census joint distribution

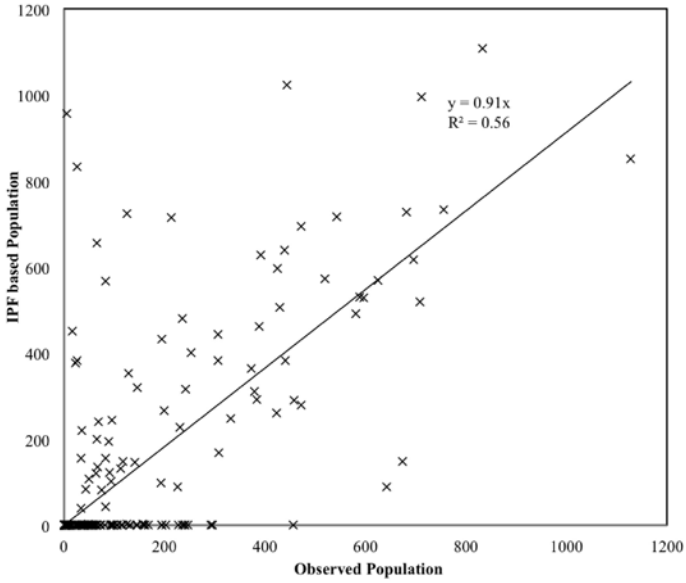


IPF with 5% sample



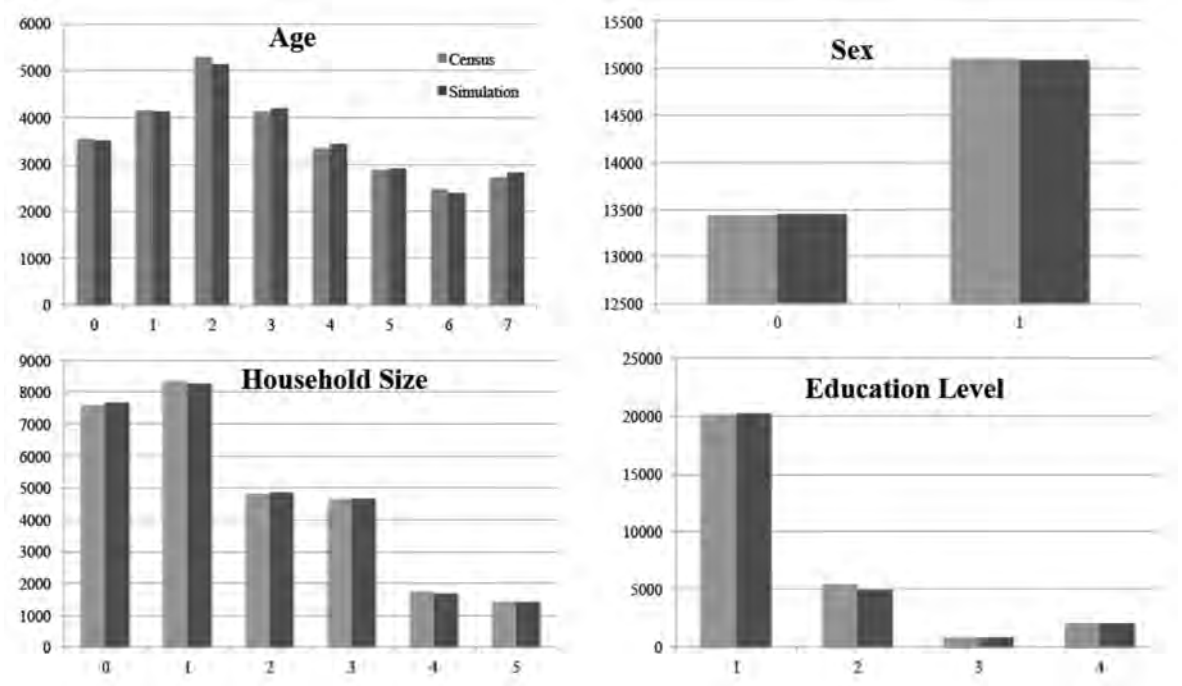
IPF with 3% sample

Results: Fit of IPF with Census joint distribution



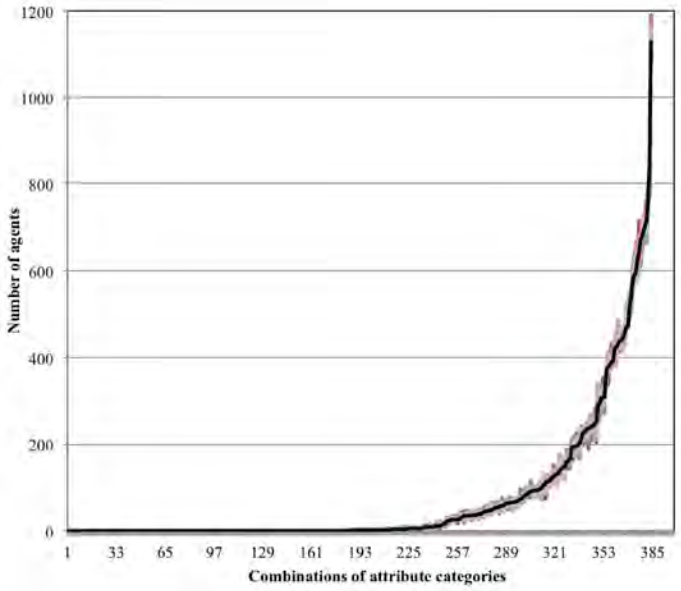
IPF with 1% sample

Results: Simulation and Census marginals



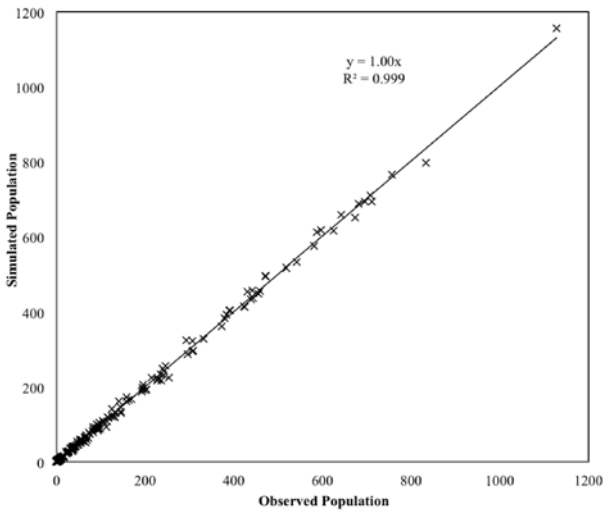
Using full-conditionals (*FullCond*)

Results: Simulation and Census joint dist.

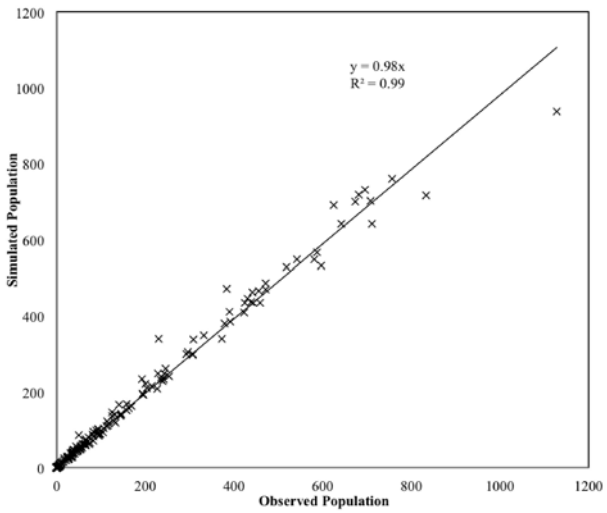


20 runs based on *FullCond* with real population superimposed

Results: Fit of Simulation with Census joint dist.

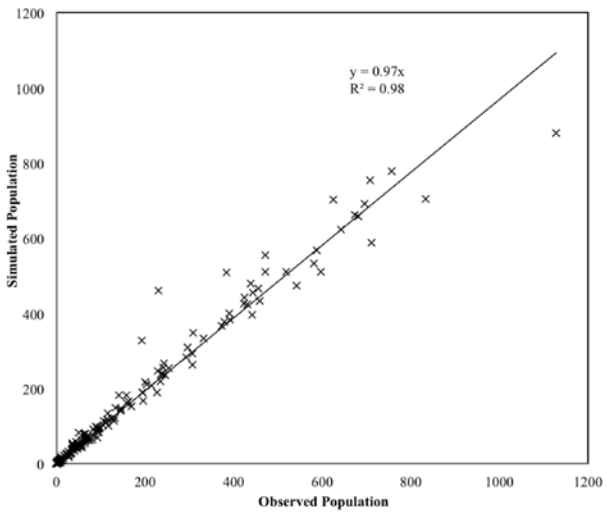


FullCond

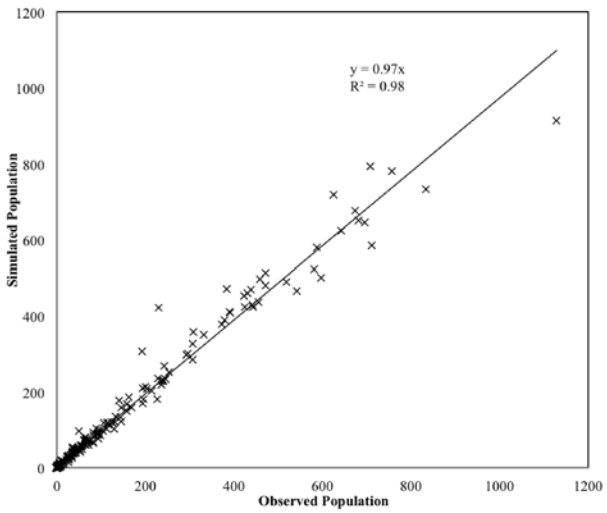


Partial_1 (Sex missing in 1 conditional)

Results: Fit of Simulation with Census joint dist.



Partial_2 (Sex missing in 2 conditionals)



Partial_3 (Sex missing in all conditional)

Comparison: Standard Root Mean Square Error

$$SRSME = \frac{\left[\sum_{i=1}^{m} \cdots \sum_{j=1}^{n} (R_{i\ldots j} - T_{i\ldots j})^2 / N\right]^{1/2}}{\sum_{i=1}^{m} \cdots \sum_{j=1}^{n} (T_{i\ldots j}) / N}$$

Comparison: Standard Root Mean Square Error

$$SRSME = \frac{\left[\sum_{i=1}^{m} \cdots \sum_{j=1}^{n} (R_{i\ldots j} - T_{i\ldots j})^2 / N\right]^{1/2}}{\sum_{i=1}^{m} \cdots \sum_{j=1}^{n} (T_{i\ldots j}) / N}$$

Input	IPF	Simulation
20%*Sample*	0.853	-
10%*Sample*	0.928	-
5%*Sample*	1.020	-
3%*Sample*	1.160	-
1%*Sample*	1.730	-
FullCond	-	0.130
Partial_1	-	0.240
Partial_2	-	0.340
Partial_3	-	0.350

Comparison: Standard Root Mean Square Error

$$SRSME = \frac{\left[\sum_{i=1}^{m} \cdots \sum_{j=1}^{n} (R_{i\ldots j} - T_{i\ldots j})^2 / N\right]^{1/2}}{\sum_{i=1}^{m} \cdots \sum_{j=1}^{n} (T_{i\ldots j}) / N}$$

Input	IPF	Simulation
20%*Sample*	0.853	-
10%*Sample*	0.928	-
5%*Sample*	1.020	-
3%*Sample*	1.160	-
1%*Sample*	1.730	-
FullCond	-	0.130
Partial_1	-	0.240
Partial_2	-	0.340
Partial_3	-	0.350

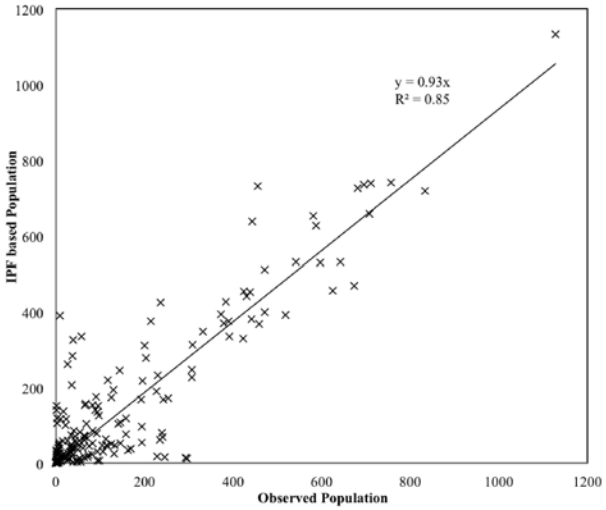
Comparison: Standard Root Mean Square Error

$$SRSME = \frac{\left[\sum_{i=1}^{m} \cdots \sum_{j=1}^{n} (R_{i\ldots j} - T_{i\ldots j})^2 / N\right]^{1/2}}{\sum_{i=1}^{m} \cdots \sum_{j=1}^{n} (T_{i\ldots j}) / N}$$

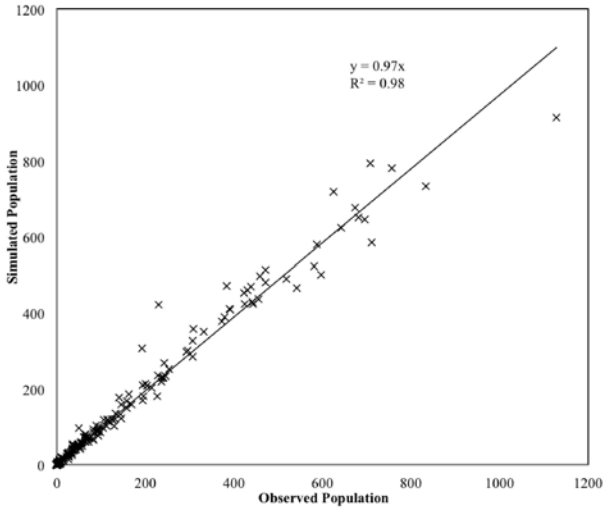
Input	IPF	Simulation
20%*Sample*	0.853	-
10%*Sample*	0.928	-
5%*Sample*	1.020	-
3%*Sample*	1.160	-
1%*Sample*	1.730	-
FullCond	-	0.130
Partial_1	-	0.240
Partial_2	-	0.340
Partial_3	-	0.350

- For Marginals only, both methods give the same fit

Best case IPF and worst case Simulation

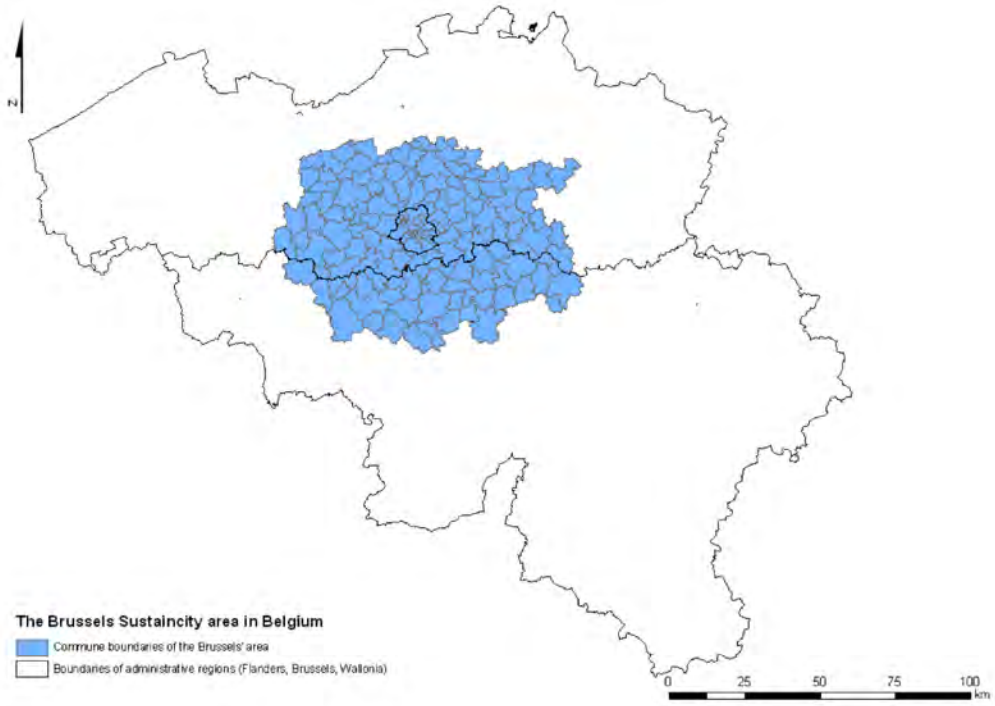


IPF with 20% sample



Partial_4 (Sex missing from all the conditionals)

Back to Brussels case study



Brussels case study

- Data sources (extremely limited)
  - Incomplete conditionals of households and persons (Census 2001)
  - Travel survey of households and individuals (MOBEL 1999)
    - **3063 observations (0.2%)**
- Synthetic household attributes
  - Size, children, workers, cars, income, university education, dwelling type, sector

Back to original problem

Brussels case study

- Data sources (extremely limited)
  - Incomplete conditionals of households and persons (Census 2001)
  - Travel survey of households and individuals (MOBEL 1999)
    - **3063 observations (0.2%)**
- Synthetic household attributes
  - Size, children, workers, cars, income, university education, dwelling type, sector

- Data Preparation
  - Aggregation
    - Spatial
    - Categorical
  - Model based conditionals (Logit)
    - Income, univ edu, cars, and dwelling type

Income level model (5 levels)

$$V^1_{(hh,z)} = 0$$

$$V^2_{(hh,z)} = ASC^2 + \beta^2_{zonal\_inc_z} \times zonal\_inc_z + \beta^2_{cars_{hh}} \times cars_{hh} + \beta^2_{workers_{hh}} \times workers_{hh}$$

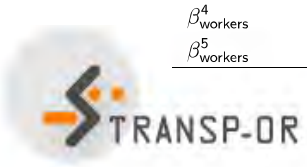
$$V^3_{(hh,z)} = ASC^3 + \beta^3_{educ_{hh}} \times educ_{hh} + \beta^3_{zonal\_inc_z} \times zonal\_inc_z + \beta^3_{cars_{hh}} \times cars_{hh} + \beta^3_{house_{hh}} \times house_{hh} + \beta^3_{workers_{hh}} \times workers_{hh}$$

$$V^4_{(hh,z)} = ASC^4 + \beta^4_{educ_{hh}} \times educ_{hh} + \beta^4_{zonal\_inc_z} \times zonal\_inc_z + \beta^4_{cars_{hh}} \times cars_{hh} + \beta^4_{house_{hh}} \times house_{hh} + \beta^4_{workers_{hh}} \times workers_{hh}$$

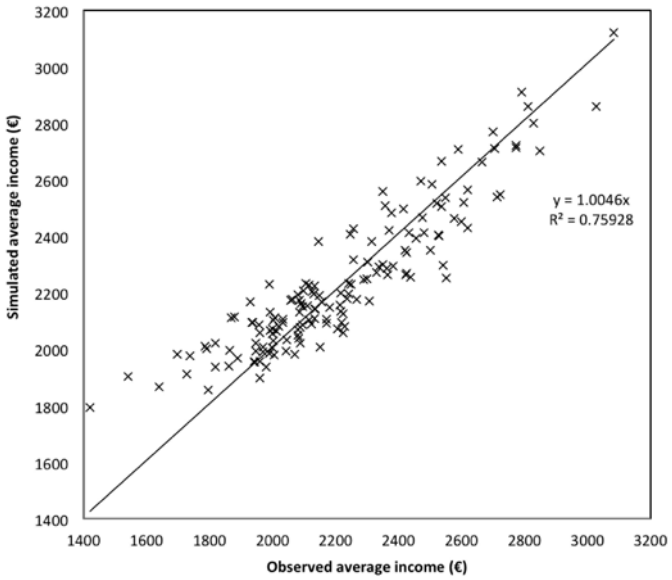
$$V^5_{(hh,z)} = ASC^5 + \beta^5_{educ_{hh}} \times educ_{hh} + \beta^5_{zonal\_inc_z} \times zonal\_inc_z + \beta^5_{cars_{hh}} \times cars_{hh} + \beta^5_{house_{hh}} \times house_{hh} + \beta^5_{workers_{hh}} \times workers_{hh}$$

Income level model

Parameter	Variable	Value	Std err	t-test
ASC^2	constant for income level 2	-0.86	0.789	-1.09
ASC^3	constant for income level 3	-4.64	0.901	-5.14
ASC^4	constant for income level 4	-8.31	1.12	-7.39
ASC^5	constant for income level 5	-10.6	1.55	-6.82
β^3_{educ}	dummy for presence of people with higher educ in the hh	0.831	0.177	4.69
β^4_{educ}	dummy for presence of people with higher educ in the hh	1.72	0.314	5.49
β^5_{educ}	dummy for presence of people with higher educ in the hh	1.92	0.656	2.93
$\beta^2_{zonal\_inc}$	average zonal income	0.0008	0.0004	1.84
$\beta^3_{zonal\_inc}$	average zonal income	0.0012	0.0005	2.55
$\beta^4_{zonal\_inc}$	average zonal income	0.0016	0.0005	3.09
$\beta^5_{zonal\_inc}$	average zonal income	0.0016	0.0006	2.47
β^2_{cars}	number of cars in the household	1.16	0.265	4.39
β^3_{cars}	number of cars in the household	1.92	0.299	6.41
β^4_{cars}	number of cars in the household	2.33	0.341	6.83
β^5_{cars}	number of cars in the household	3.2	0.466	6.87
β^3_{house}	dummy for dwelling being a house	0.45	0.193	2.34
β^4_{house}	dummy for dwelling being a house	0.485	0.294	1.65
β^5_{house}	dummy for dwelling being a house	0.485	0.294	1.65
$\beta^2_{workers}$	number of workers in the household	1.14	0.277	4.11
$\beta^3_{workers}$	number of workers in the household	2.22	0.295	7.53
$\beta^4_{workers}$	number of workers in the household	2.46	0.345	7.13
$\beta^5_{workers}$	number of workers in the household	1.74	0.428	4.07

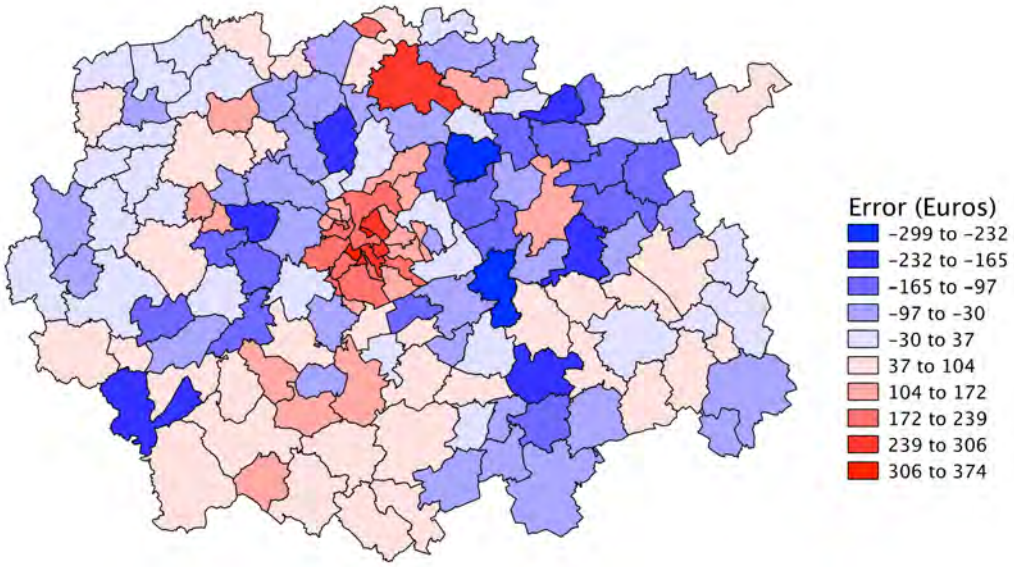


Results: Brussels case study



Fit between simulation based and observed average commune-level income

Results: Brussels case study



Spatial distribution of error in average income

- More zonal level demographic statistics are required to further decrease the error

Concluding remarks

- From single solution optimization problem to sampling from joint distribution
  - Output of microsimulation models

$$O = \int_{p_{syn}} microsim(p_{syn})\, dp_{syn}.$$

- Focus on reproducing not just marginals, but the whole joint distribution
- Heterogeneous not cloned population
- Population synthesis as part of microsimulation
  - Sensitivity analysis in a coherent way
- Separation of data preparation from agent generation
  - Data, models, assumptions

Concluding remarks

- Mix of sampling process can be utilized based on the situation
- Works both for continuous and discrete or mixture of conditionals
- Computationally efficient and scalable
  - Clean and simple
- Issue of inconsistency
  - Open research question [Buuren, 2007][Chen et al., 2011]
- Use of new and unconventional data
  - WiFi network (Pedestrian movement)
  - Online check-in / social media
- Resource and Agents association
  - from bi-partite to k-partite graph [Anderson et al., 2014]

Bibliography I

Anderson, P., Farooq, B., Efthymiou, D., and Bierlaire, M. (2014).
Association generation in synthetic population for transportation applications: Graph-theoretic solution.
Transportation Research Record, 2429:38–50.

Beckman, R. J., Baggerly, K. A., and McKay, M. D. (1996).
Creating synthetic baseline populations.
Transportation Research Part A: Policy and Practice, 30(6):415–429.

Buuren, S. V. (2007).
Multiple imputation of discrete and continuous data by fully conditional specification.
Statistical Methods in Medical Research.

Bibliography II

Chen, S.-H., Ip, E. H., and Wang, Y. J. (2011).
Gibbs ensembles for nearly compatible and incompatible conditional models.
Comput. Stat. Data Anal., 55(4):1760–1769.

Farooq, B., Bierlaire, M., Hurtubia, R., and Flötteröd, G. (2013).
Simulation based population synthesis.
Transportation Research Part B: Methodological, 58:243–263.

Farooq, B., Hurtubia, R., and Bierlaire, M. (2015).
Simulation based generation of a synthetic population for brussels.
In Bierlaire, M., de Palma, A., Hurtubia, R., and Waddell, P., editors, *Integrated Transport and Land Use Modeling for Sustainable Cities*, pages 95–112. EPFL Press.
ISBN:978-2-940222-72-8.

Bibliography III


Geman, S. and Geman, D. (1984).
Stochastic relaxation, gibbs distributions, and the bayesian restoration of images.
Pattern Analysis and Machine Intelligence, IEEE Transactions on, PAMI-6(6):721 –741.

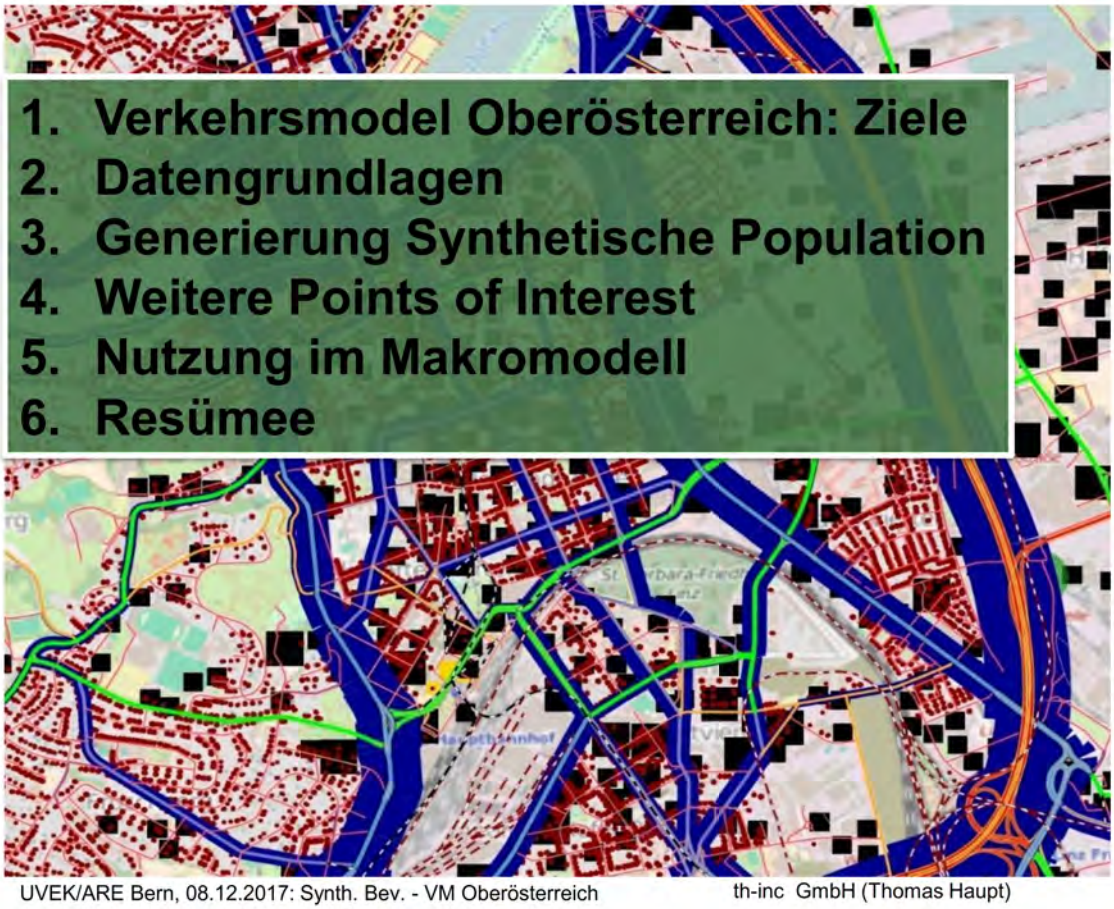

Hubert, J. P. and Toint, P. L. (2002).
La mobilite quotidienne des belges.
Mobilite et Transports, 1.


Williamson, P., Birkin, M., and Rees, P. H. (1998).
The estimation of population microdata by using data from small area statistics and samples of anonymised records.
Environment and Planning A, 30(5):785–816.

Aufbau und Anwendung einer synthetischen Bevölkerung im Verkehrsmodell Oberösterreich

Thomas Haupt
th-inc GmbH/ Senozon Deutschland GmbH

Inhalt



Ziele des Verkehrsmodells Oberösterreich

1. **Umfassende und integrierte Datengrundlage für verkehrsplanerische Untersuchungen im IV und ÖV**
2. **Simulation und Prognose von verkehrlichen Maßnahmen**
3. **Wirkungsanalyse von demographischen, strukturellen und feinräumigen Veränderungen und Bauvorhaben**
4. **Grundlage und Werkzeug für Einnahmeberechnung und Ausschreibungen von ÖV-Leistungen**
5. **Unterstützung, Erleichterung und qualitative Verbesserung von Bearbeitungs- und Entscheidungsprozessen im Land und im Austausch mit den Partnern**
6. **Einbezug der Verkehrstelematik und Verkehrsinformation in die Verkehrsmodellierung (future use)**
7. **Unterstützung von Multi- und Intermodalität**

Anforderungen an das Verkehrsmodell

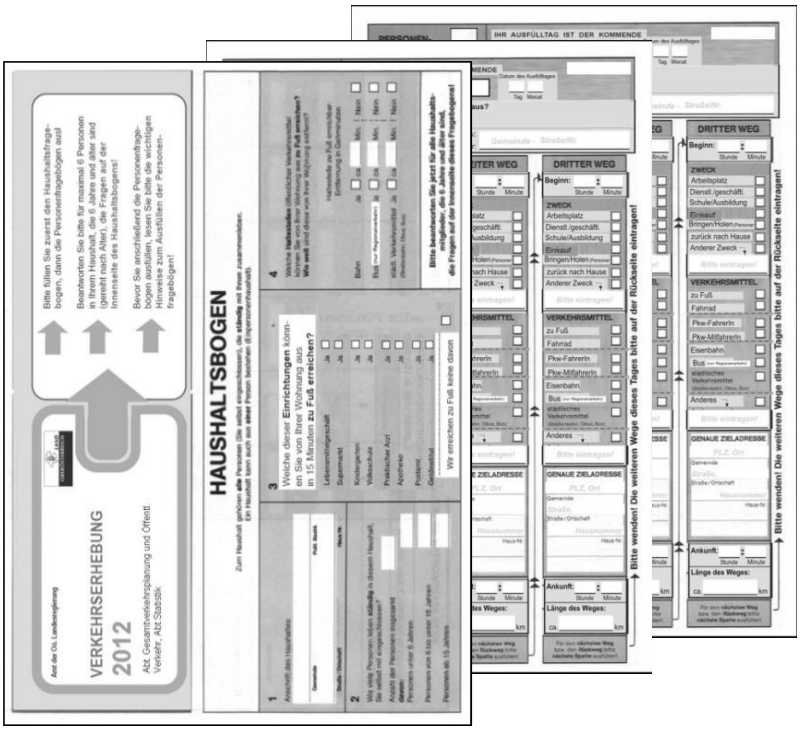
1. **Easy to use**
2. **E**xtract **T**ransform **L**oad Prozess (ETL)
3. **Transparente Dokumentation**
4. **Korrektheit und Robustheit**
5. **Konsistenz des Datenmodells**
6. **Korrekte Projektion**

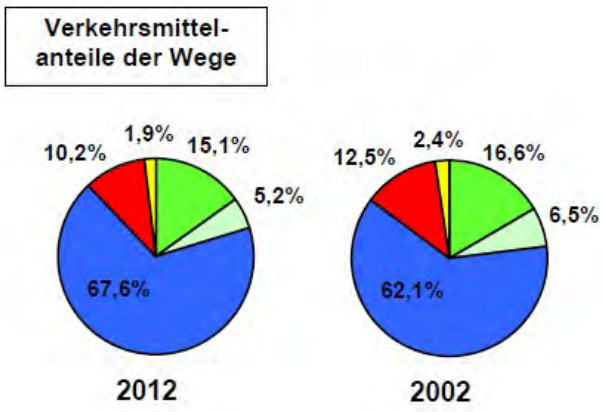
Datengrundlagen

1. **Detaillegetreue digitale Straßennetze**
2. **Digitale Fahrplandaten**
3. **Umfangreiche HH-Befragung 2012**
4. **Digitale hochaufgelöste Gebäude und POI-Informationen**
5. **Gebiete, Kataster, 250m und 100m Raster, Adressdaten**
6. **OpenStreetMap**

Haushaltsbefragung 2012

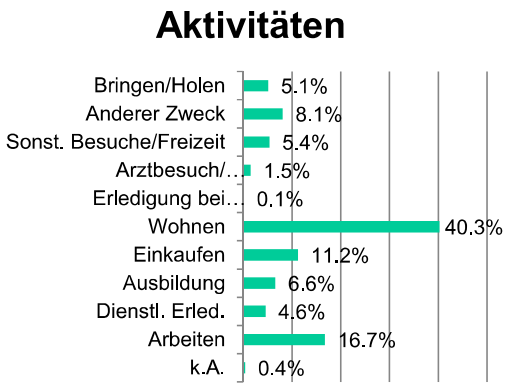
Grundlagen der Mobilität - Oö. Verkehrserhebung/Haushaltsbefragung 2012,
Abbildung des Mobilitätsverhalten mit über 80T HH, 200T Personen und 600T Wege

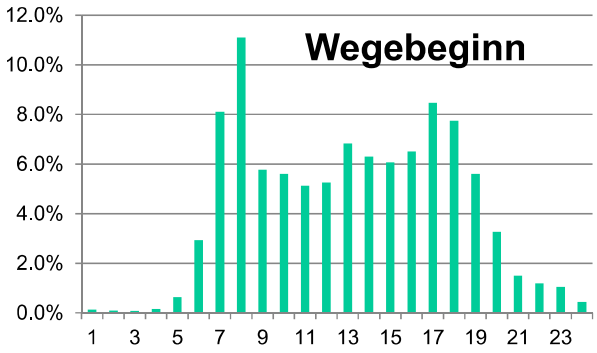


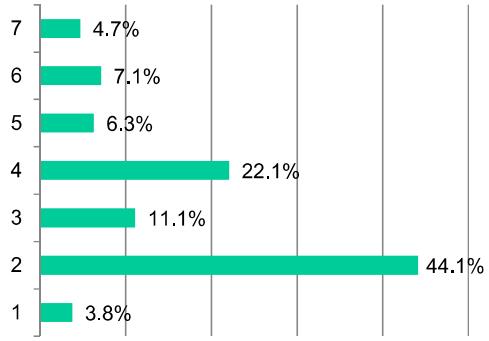


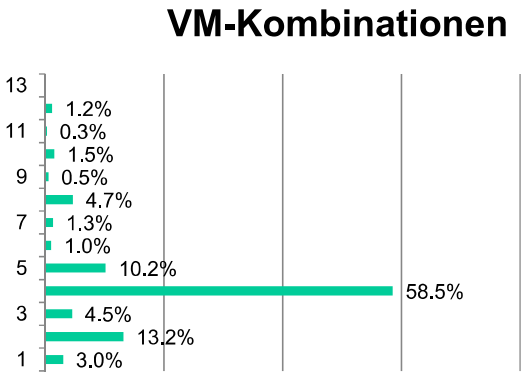
Quelle: Pfeiffer/Haupt: AWS PTV 10/2015

HH-Befragung Oberösterreich 2012: Aktivitäten und VM-Wahl









Quelle: Pfeiffer/Haupt: AWS PTV 10/2015: Verkehrsmodell Oberösterreich

Vorgehen im Verkehrsmodell Oö: Strukturdaten und HH-Befragung

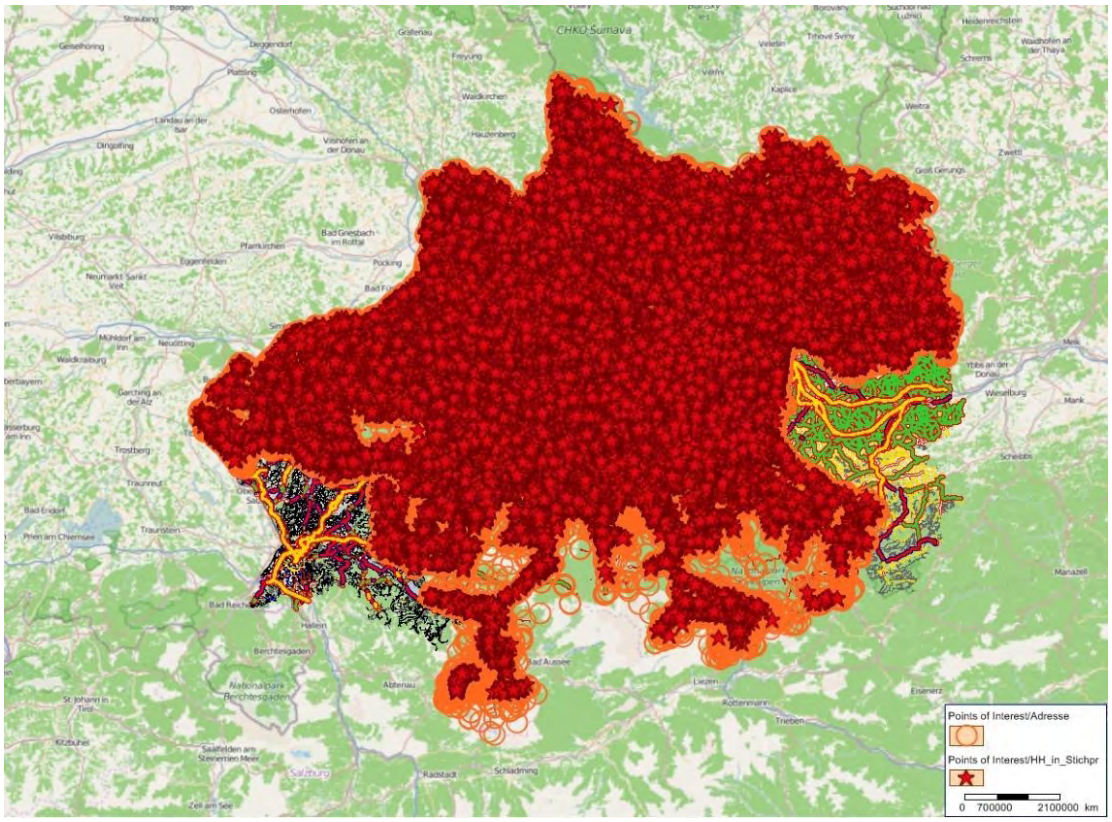
Ideen:

1. **HH-Befragung randomisiert in VISUM darstellen**
2. **Alle Einwohner als POI mit statistisch korrekten individuellen Merkmalen**
3. **Synthetische Bevölkerung erzeugen**
4. **Einwohner-POI zu verhaltenshomogen Gruppen aggregieren**
5. **Alle Strukturdaten für Verkehrszellen aus POI ableiten (Verschneiden nutzen)**

Ausgangspunkt: 365T Adressdatenpunkte

inkl. 141 Altenheime, 29 Studentenwohnheim und 7 Justizanstalten

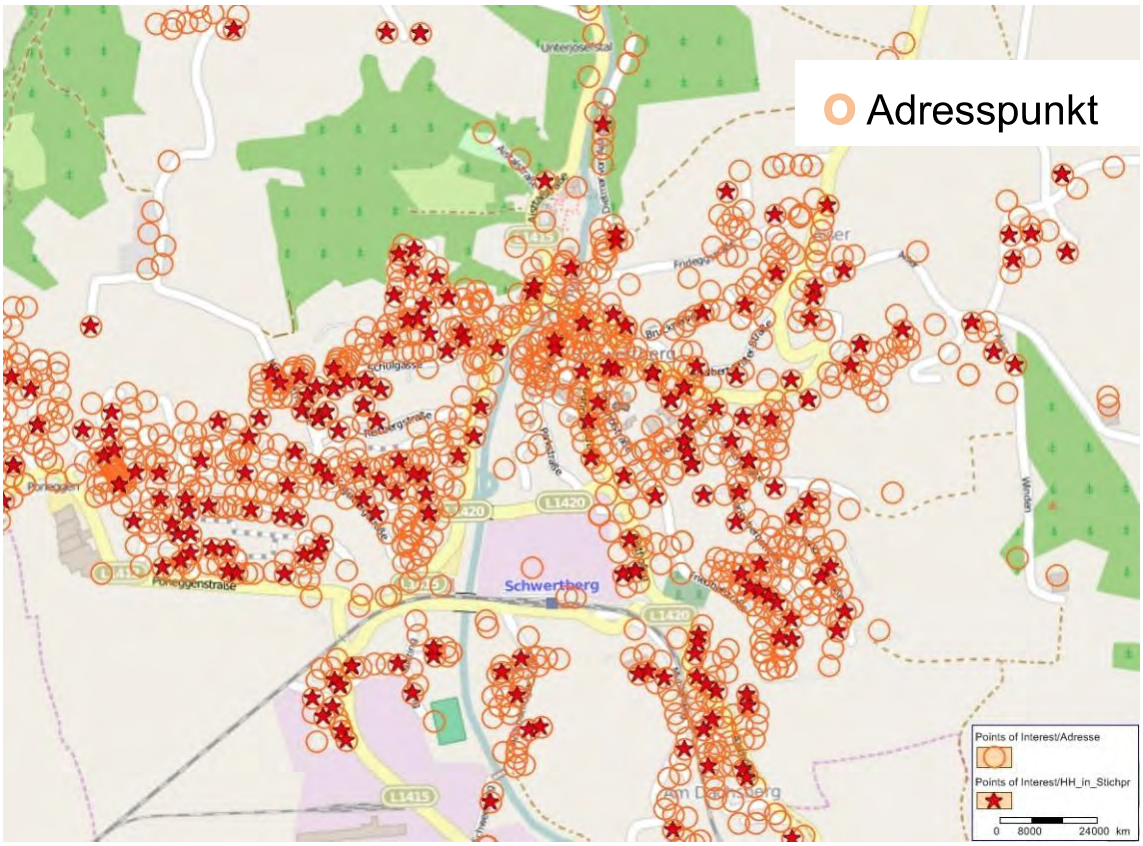
HH-Befragung: Antwortstichprobe



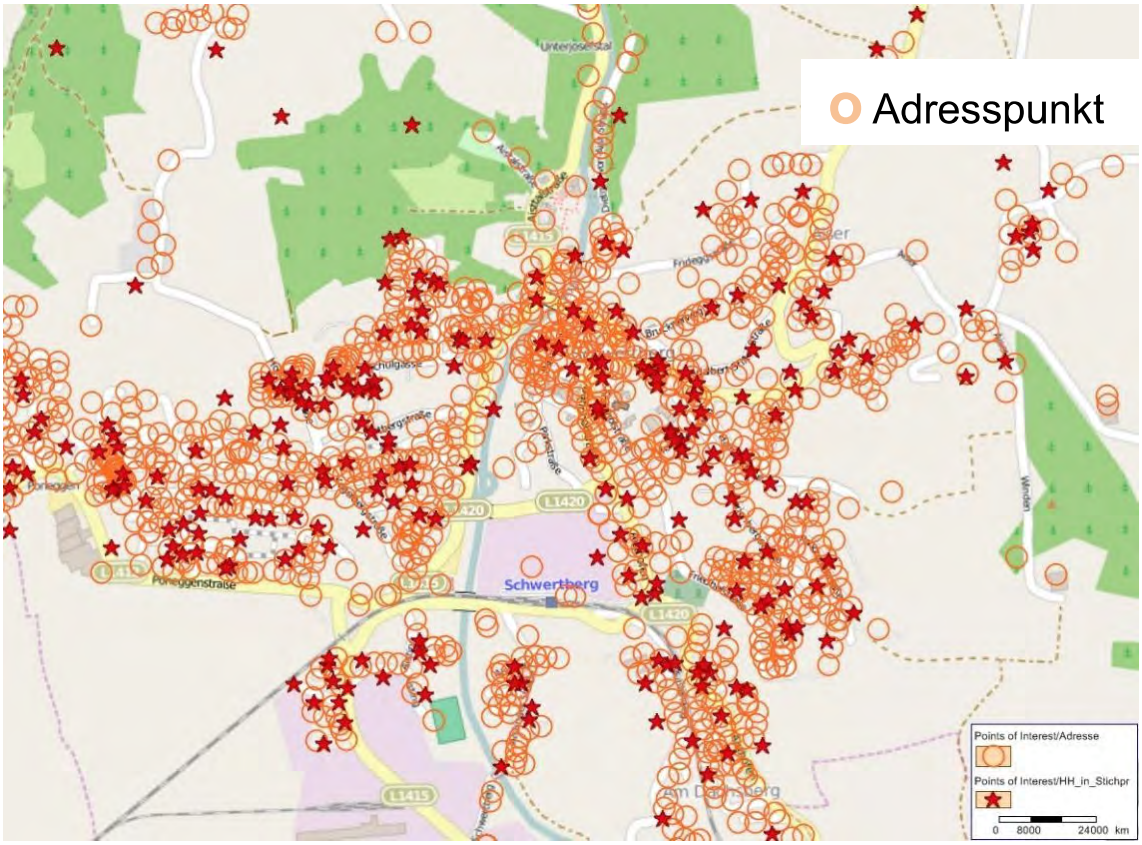
HH-Befragung 2012:

- 83T HH
- 200T von 1,4 Mio EW
- 600T Wege

Verkehrsmodell Oö: HH-Befragung Stichprobe



Verkehrsmodell Oö: HH-Befragung **Randomisiert**



Fitting: Adresspunkte in Zonen/Verkehrszellen & Raster

Ca. 1200 Zonen/Verkehrszellen

Ca. 56000 250m-Raster

Basisauswertungen

<u>aus der Statistikstelle des Landes:</u>

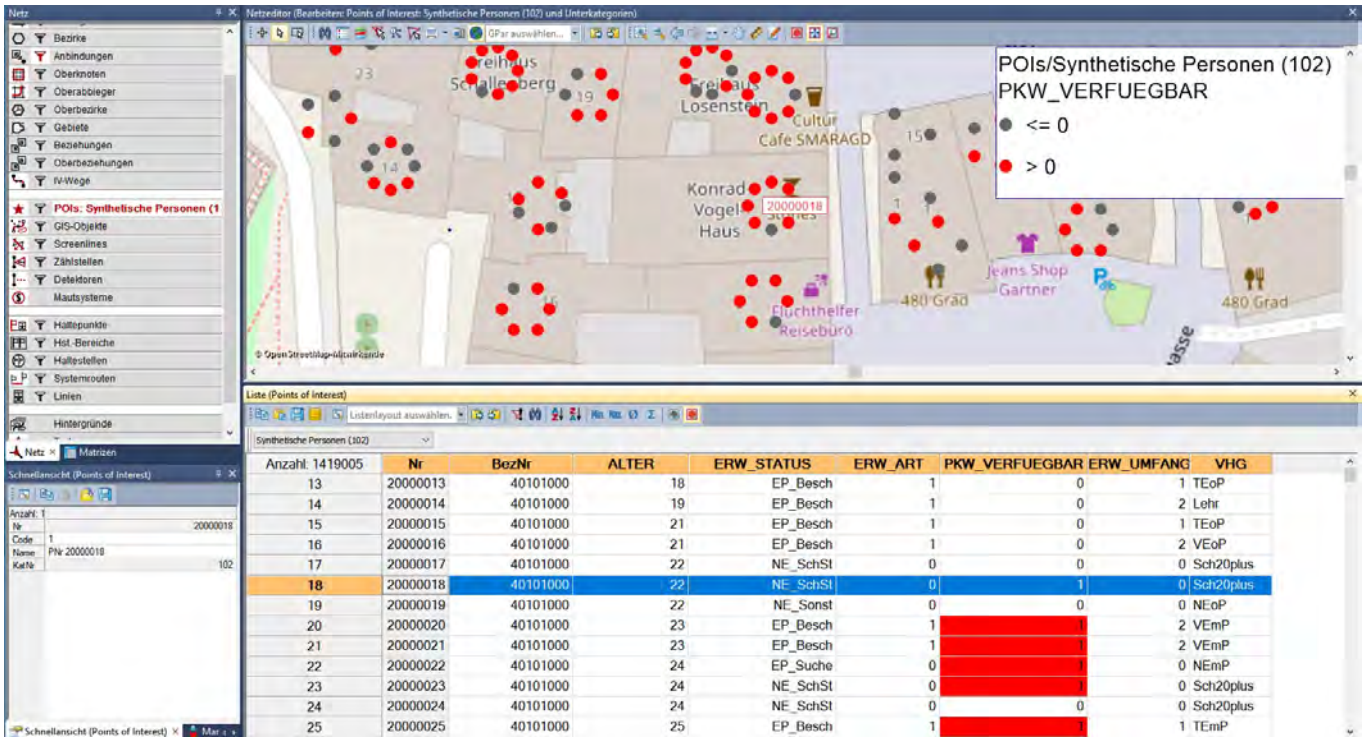
Kreuzklassifikation der Wohnbevölkerung

1) **Raster (56000)**
2) **Zählsprengel (ca. 1200)**
3) **Geschlecht (2)**
4) **Alter**
5) **Erwerbsstatus (6 Ausprägungen gemäß Statistik Austria)**
6) **Erwerbsart (3 Ausprägungen, keine, unselbständig, selbständig)**

<u>aus der Haushaltsbefragung</u>

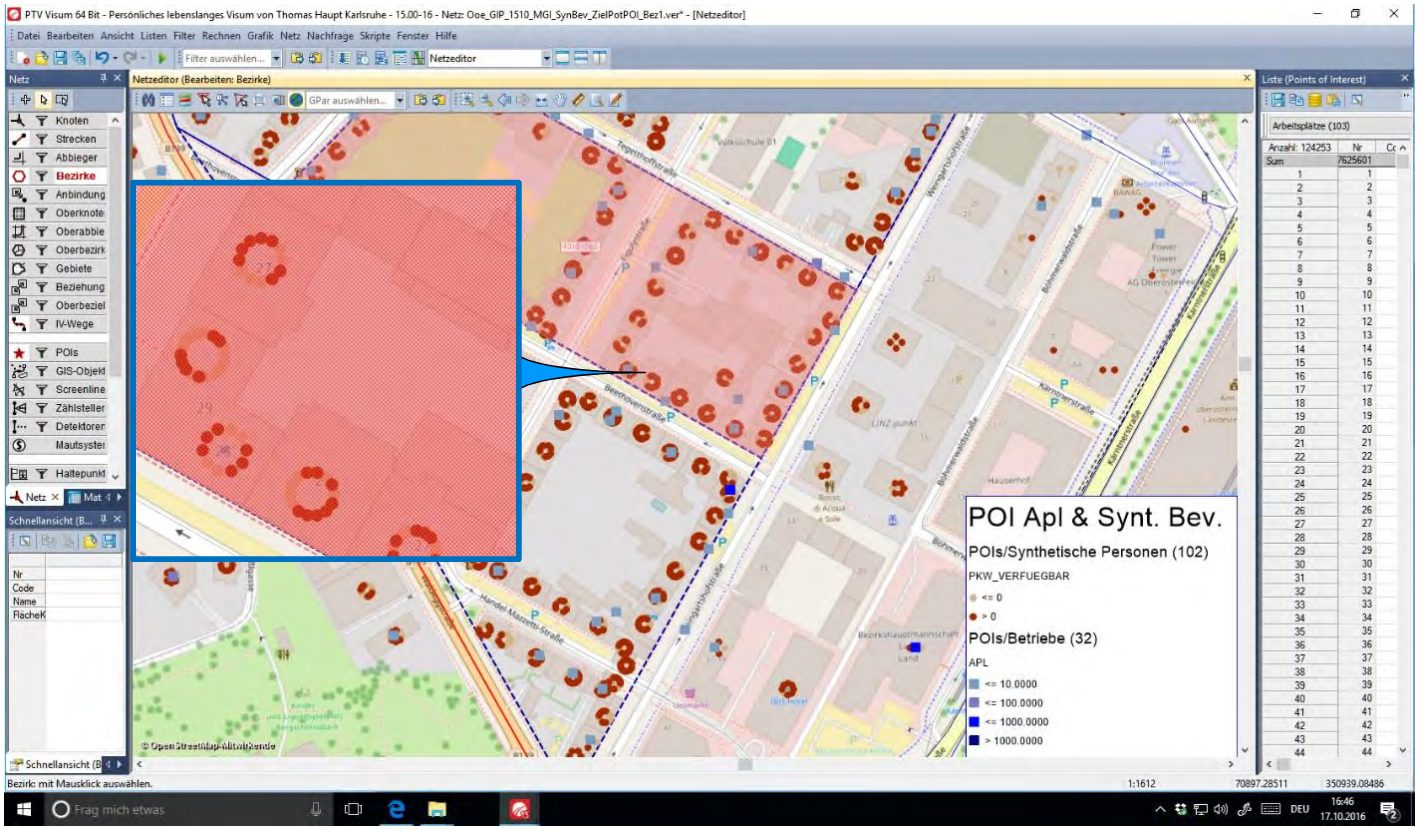
1) **Pkw-Verfügbarkeit (0 oder 1)**
2) **Erwerbstätigkeit (keine, Teilzeit, Vollzeit)**

Synth. Bev.: 1,4 Mio Einwohner als POI

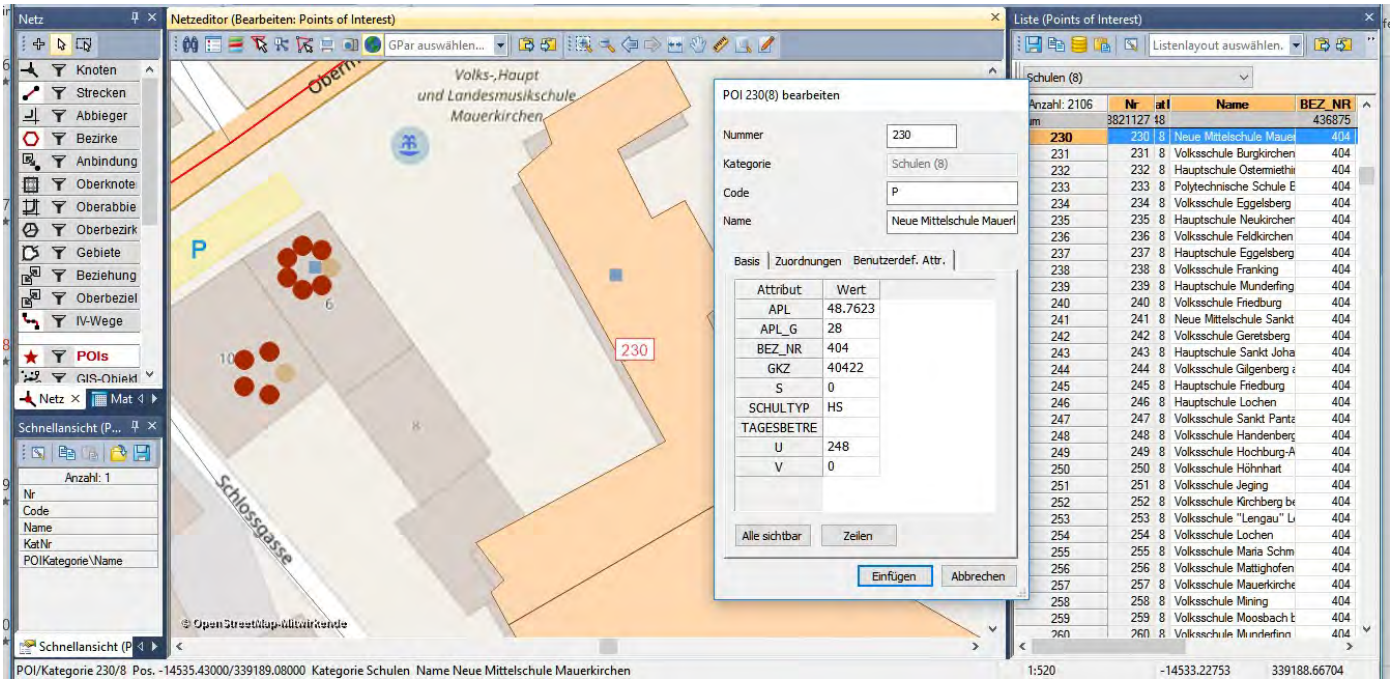


Nr	BezNr	ALTER	ERW_STATUS	ERW_ART	PKW_VERFUEGBAR	ERW_UMFANG	VHG
13	20000013	40101000	18	EP_Besch	1	0	1 TEoP
14	20000014	40101000	19	EP_Besch	1	0	2 Lehr
15	20000015	40101000	21	EP_Besch	1	0	1 TEoP
16	20000016	40101000	21	EP_Besch	1	0	2 VEoP
17	20000017	40101000	22	NE_SchSt	0	0	0 Sch20plus
18	20000018	40101000	22	NE_SchSt	0	1	0 Sch20plus
19	20000019	40101000	22	NE_Sonst	0	0	0 NEoP
20	20000020	40101000	23	EP_Besch	1	1	2 VEmP
21	20000021	40101000	23	EP_Besch	1	1	2 VEmP
22	20000022	40101000	24	EP_Suche	0	1	0 NEmP
23	20000023	40101000	24	NE_SchSt	0	1	0 Sch20plus
24	20000024	40101000	24	NE_SchSt	0	0	0 Sch20plus
25	20000025	40101000	25	EP_Besch	1	1	1 TEmP

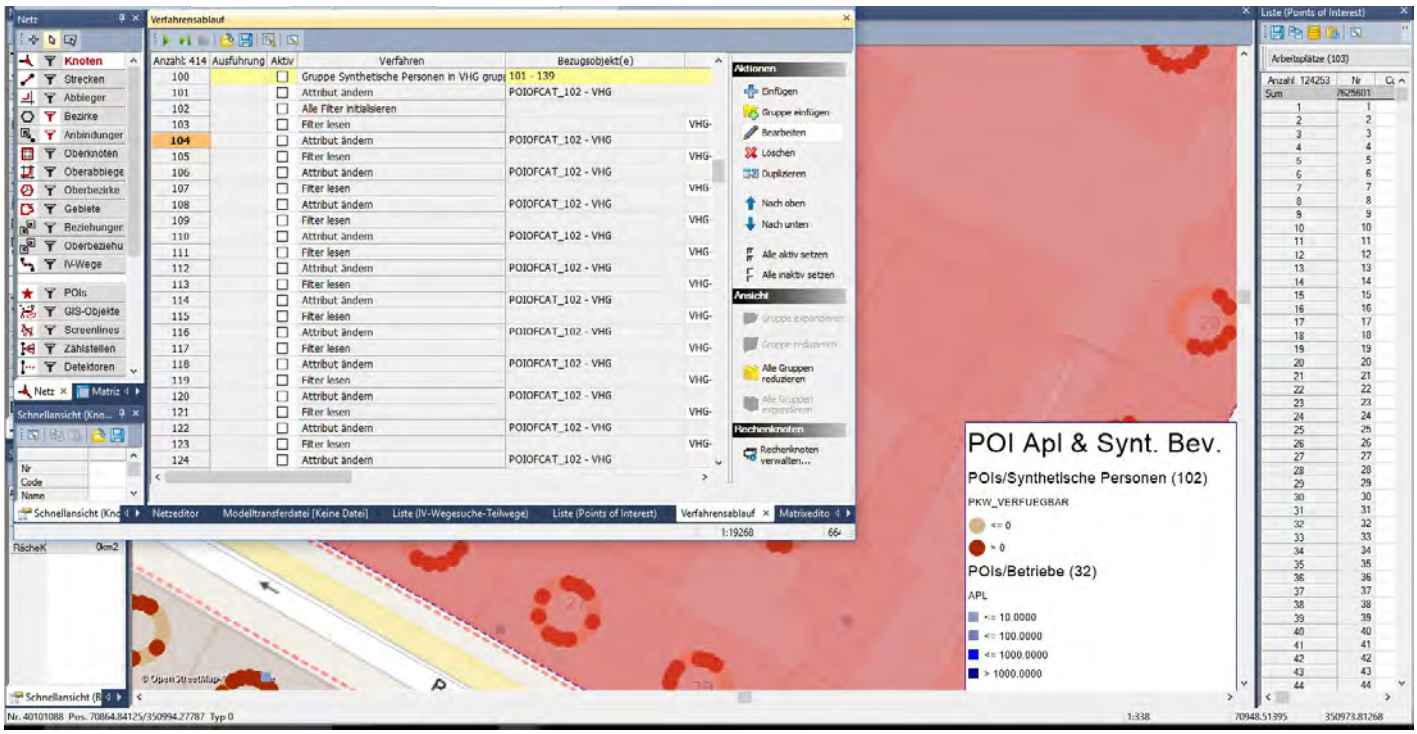
Bevölkerung als POI



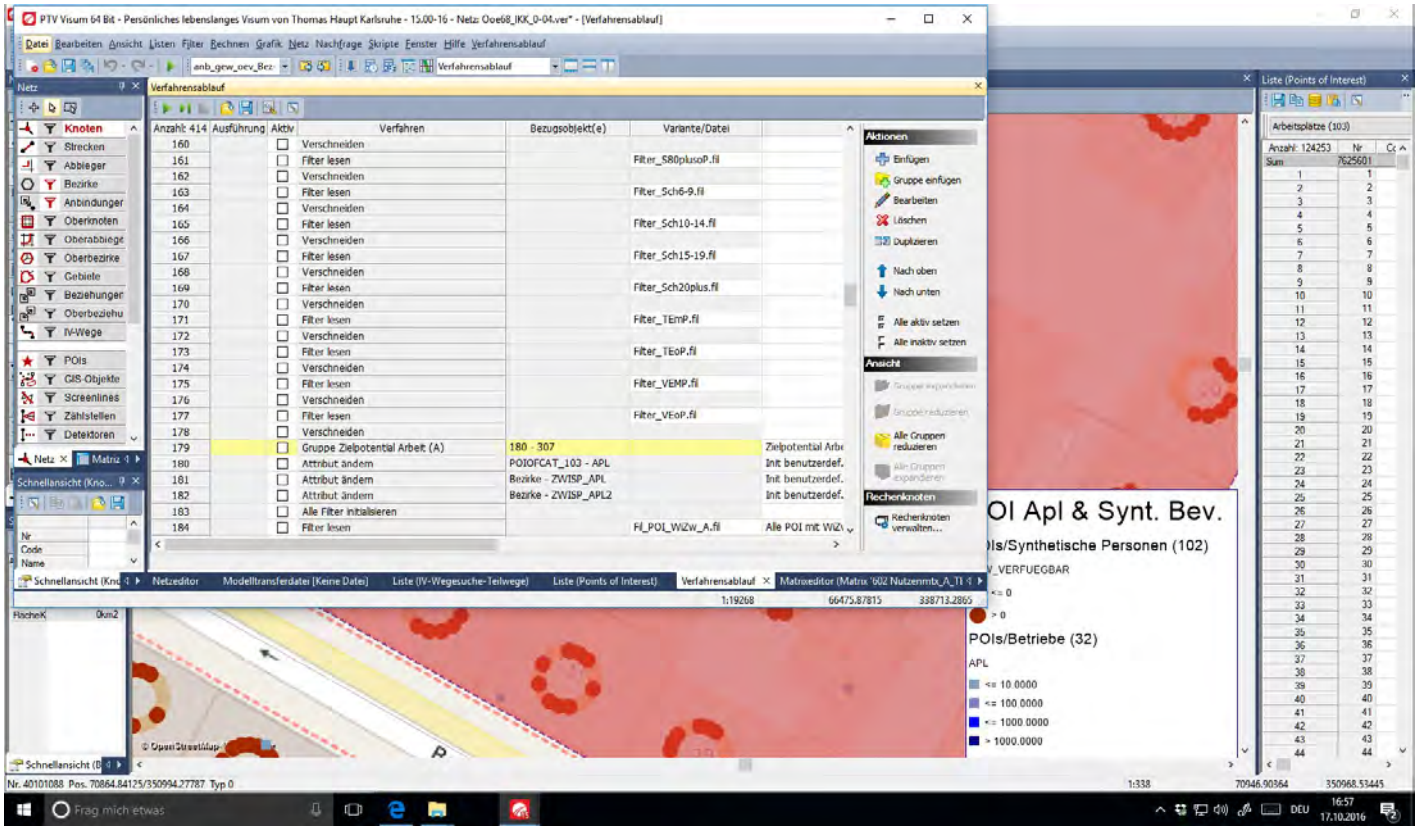
Strukturdaten je Schule



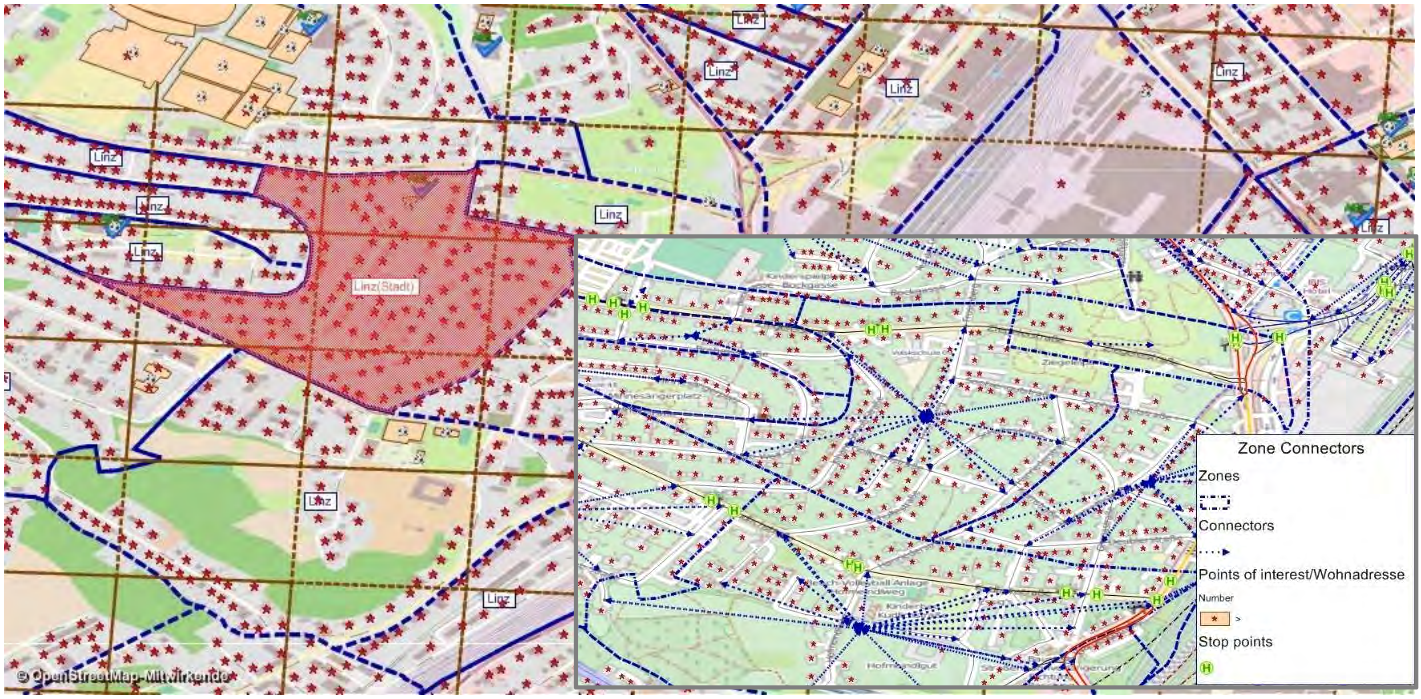
Verhaltenshomogene Gruppen zuordnen



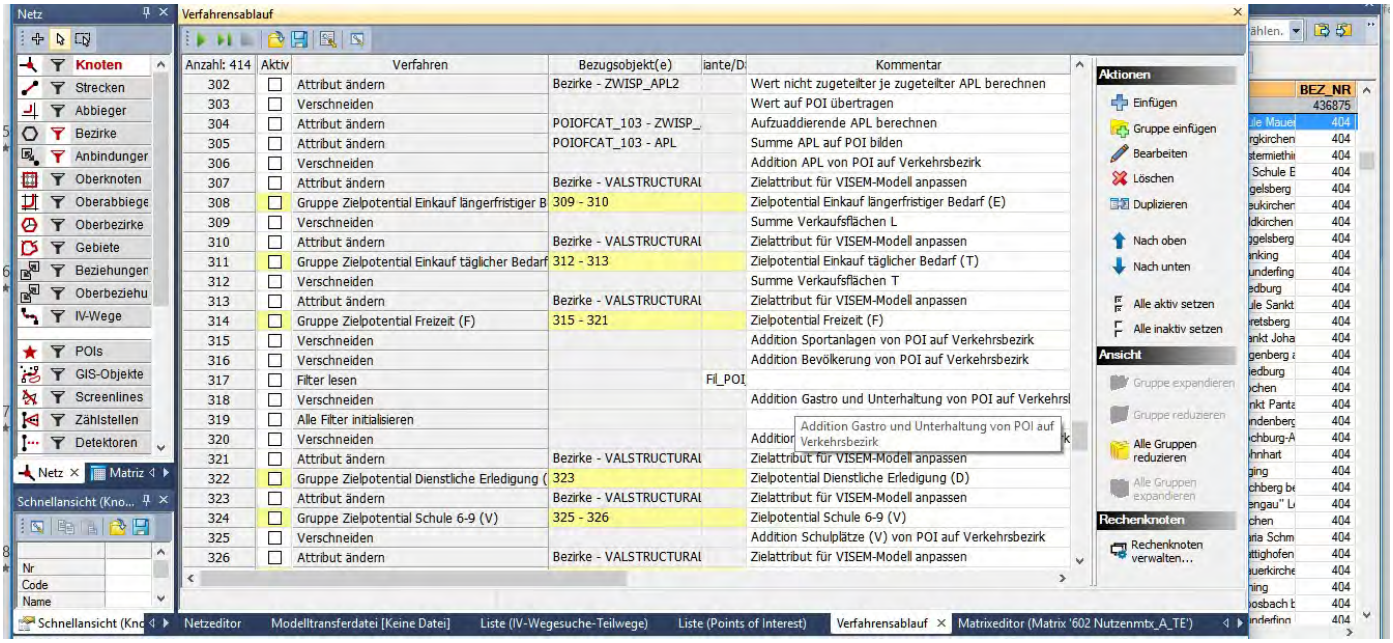
Strukturwerte für Verkehrszellen ermitteln



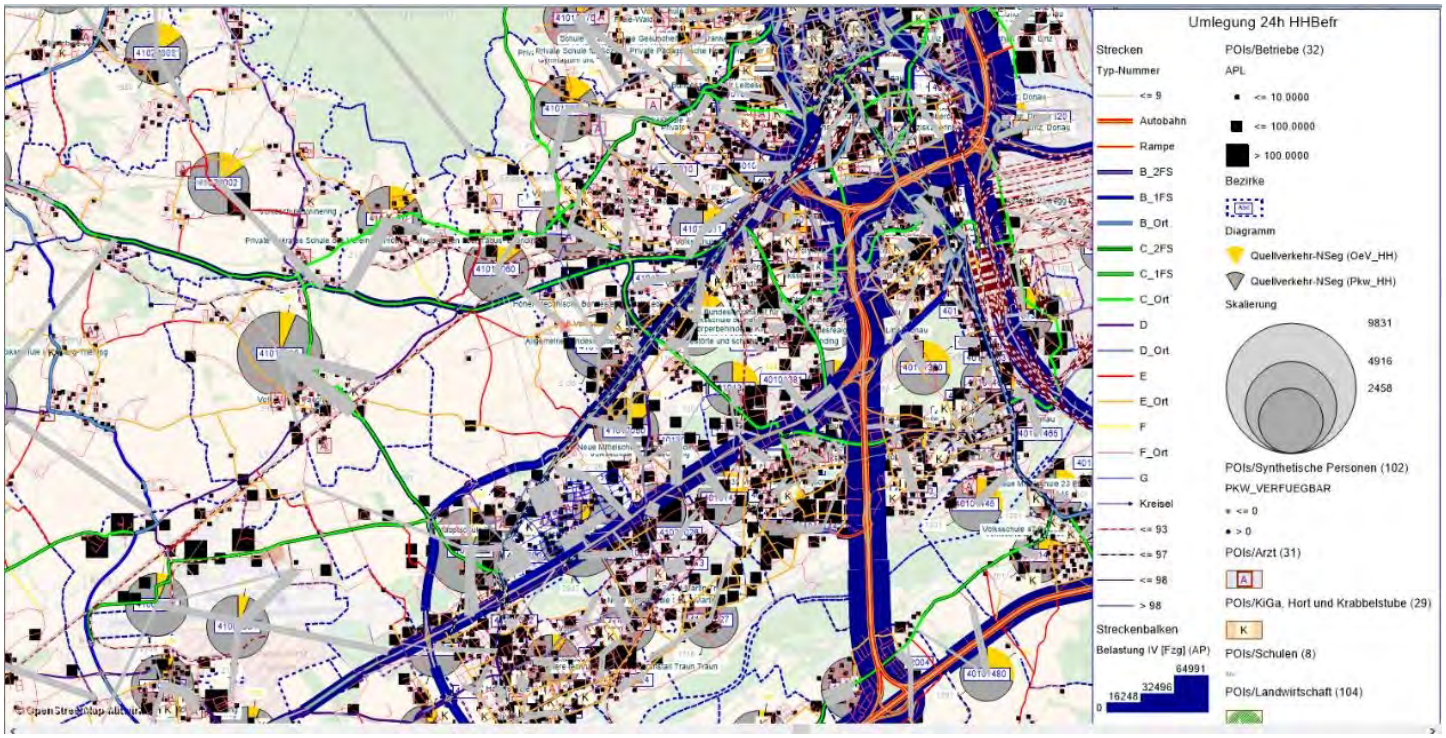
Anbindungen und Gewichte aus POI erzeugen



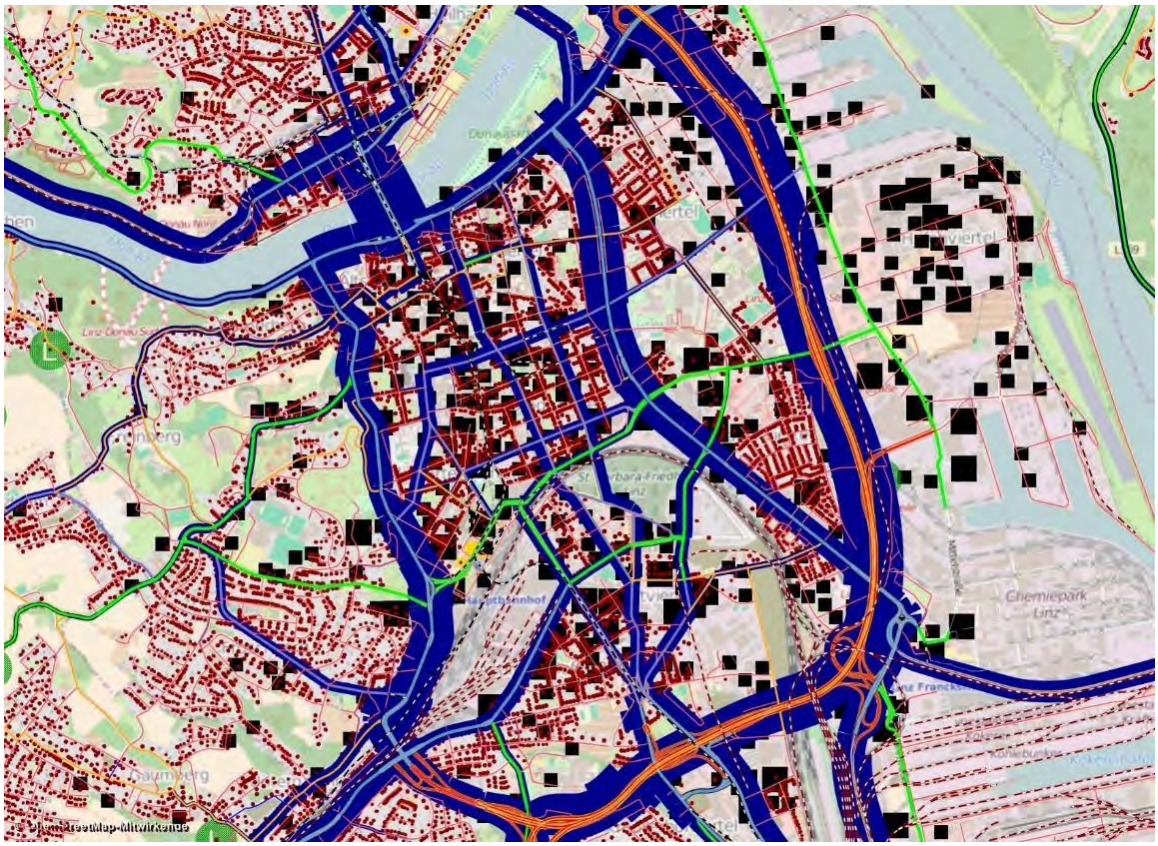
Schulplätze auf Bezirke addieren (durch Verschneiden)



Verkehrsmodell Oö: VM-Anteile und Arbeitsplätze



Einwohner- und Arbeitsplatz-POI



Resümee

1. Geocodierung und Randomisierung der HH-Befragung
2. Synthetische Bevölkerung → maximale, zonenfreie Datenauflösung
3. POI und Verschneidung → bottom up Datenversorgung
4. Hybrid (aggreg. + disaggr.) → ein Datensatz, verschiedene SW
5. Anbindungen → automatisch und detailliert

<u>Weitere Hinweise:</u>

- Digitale Geographie → Exakte geographische Projektion
- Extract, Transform, Load → Automatisierter, wiederholbarer Prozess
- ÖV-Integration → autom. Prozess in VISUM
- OSM-Nutzung und Luftbild → Infosystem für den/die Planer/in

Kontakt:

Thomas.Haupt@th-inc.de

Thomas.Haupt@senozon.com

Vielen Dank für Ihre Aufmerksamkeit!

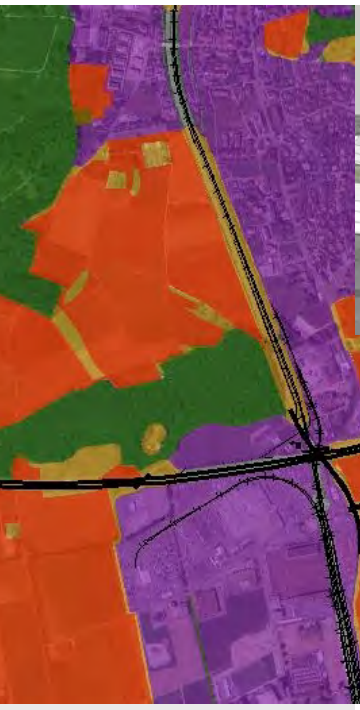
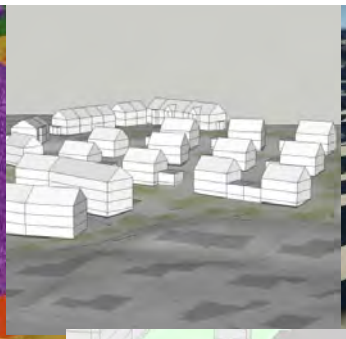
Synthetische Populationen aus FaLC-sim für das Nationale Personenverkehrsmodell NPVM

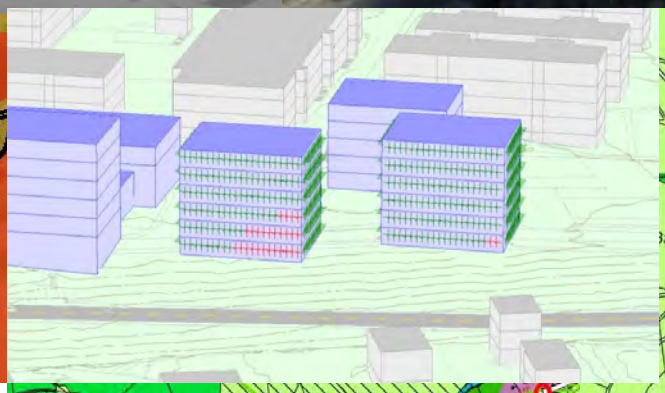
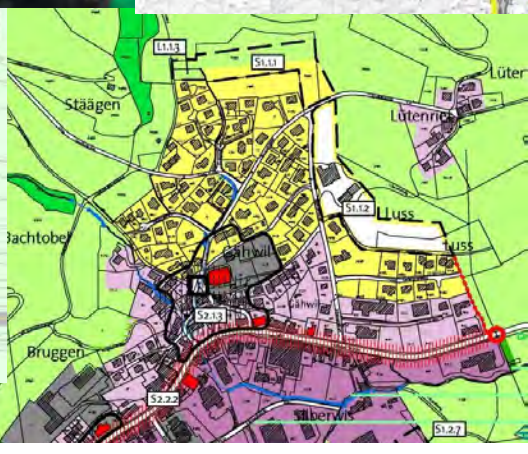
Strittmatter Partner AG

Tagung: Synthetische Populationen für die Politikberatung
8. Dezember 2017, Bern
Balz Bodenmann, Strittmatter Partner AG

Strittmatter Partner AG ...

... ein Planungsbüro !

Synthetische Populationen …

… nutzen wir in unserem Alltag !

- Bevölkerung und Arbeitsplätze – heute und morgen
  - Schulraumplanung / Altersheimplanung
  - Steuereinnahmen
  - Nachfrage nach Wohn- und Arbeitsflächen
  - Prüfung Visualisierung plausibler Bautätigkeit
  - Planung Haltestellen im Öffentlichen Verkehr
  - Auswirkungen Nutzung Autonomer Fahrzeuge auf die Siedlungsentwicklung
- Verkehrsmodelle
  - Parkplatzplanung
  - Auswirkung Umfahrungen / Änderung Infrastruktur
  - Auswirkungen der Siedlungsentwicklung / Grossprojekte
  - Erreichbarkeit von Kunden

Synthetische Populationen …

… nutzen wir in unserem Alltag !

- Basisdaten für andere Modelle
  - Gebäudeparkmodell
    Energieverbrauch heute und morgen
    (TEP Energy)
  - Kopplung von anthropogenen und ökologischen Netzwerken für eine nachhaltige Landschafts- und Verkehrsplanung (ETH Zürich, PLUS)
  - Strukturdaten für Verkehrsmodelle
    (ARE NPVM, verschiedene Kantone)

FaLC: Idee und …

… Umsetzung.

Simulations–Tool FaLC

Zu beantwortende Fragen:

FaLC-sim

was sind die Effekte von …

Szenarien

- **Politischen Entscheiden**
  - neuen **Infrastruktur-Projekten** (Strassen, öffentlicher Verkehr)
  - Änderungen von **Steuersätzen / Anreizen**
  - Änderungen von **Gesetzen / Regulativen** (z.B. Bauzonen)

Szenarien

- **Wirtschaft**
  - **Wirtschaftskrisen** (z.B. Arbeitsplatzabbau)
  - Veränderungen in Markt-Mechanismen (insb. Immobilien)
  - Standortwahl von (sehr) **grossen Unternehmen**

Zu beantwortende Fragen:

FaLC-sim

was sind die Effekte auf …

Demographie

- **Demographie**
  - **Anzahl / Alter der Einwohner**
  - **Einkommen / Steuern**
  - **räumliche Segregationen**

Standortwahl HH

Firmographie

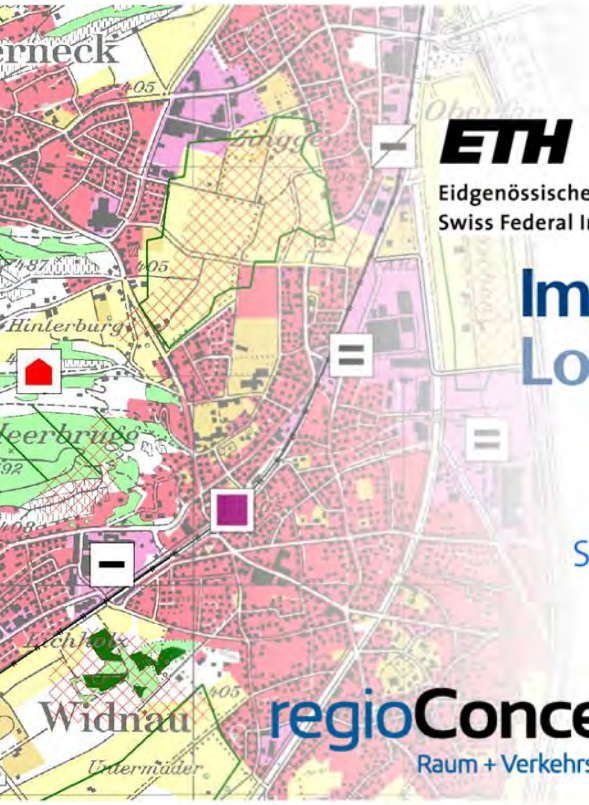
- **Firmographie**
  - **Sektoren und Grösse der Unternehmen**
  - **generierte Arbeitsplätze**
  - **generiertes Steuereinkommen**

Standortwahl Firmen

- **Gesellschaftliche, soziale und politische Ziele**
  - **Erschwinglichkeit von Land und Wohnraum**
  - **CO2-Immissionen (etc.)**
  - **Verhinderung Urban Sprawl**
  - **…**

diverse Modelle

Partner



Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Bundesamt für Raumentwicklung ARE

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Imperial College London

SBB CFF FFS

FP RE Fahrländer Partner Raumentwicklung

Strittmatter Partner AG

simrpa

regioConceptAG
Raum + Verkehrsentwicklung

ESMO

clavis:IT
winning solutions for leading companies

Siedlung | Landschaft | Verkehr | Umwelt | Recht | Wir schaffen Lebensräume.

Simulation Flächennutzung

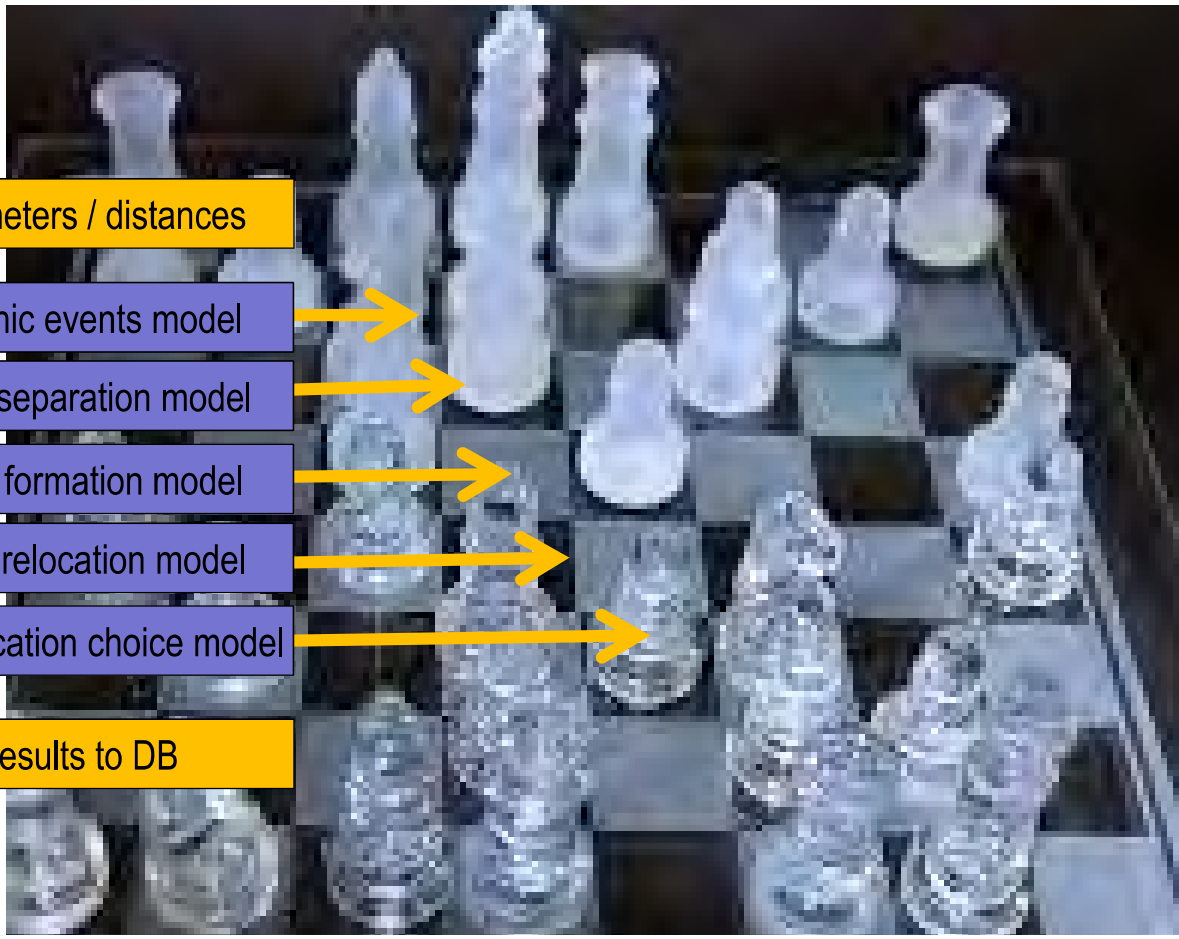
Mikrosimulation mit Agenten

- Personen
- Haushalte
- Unternehmen

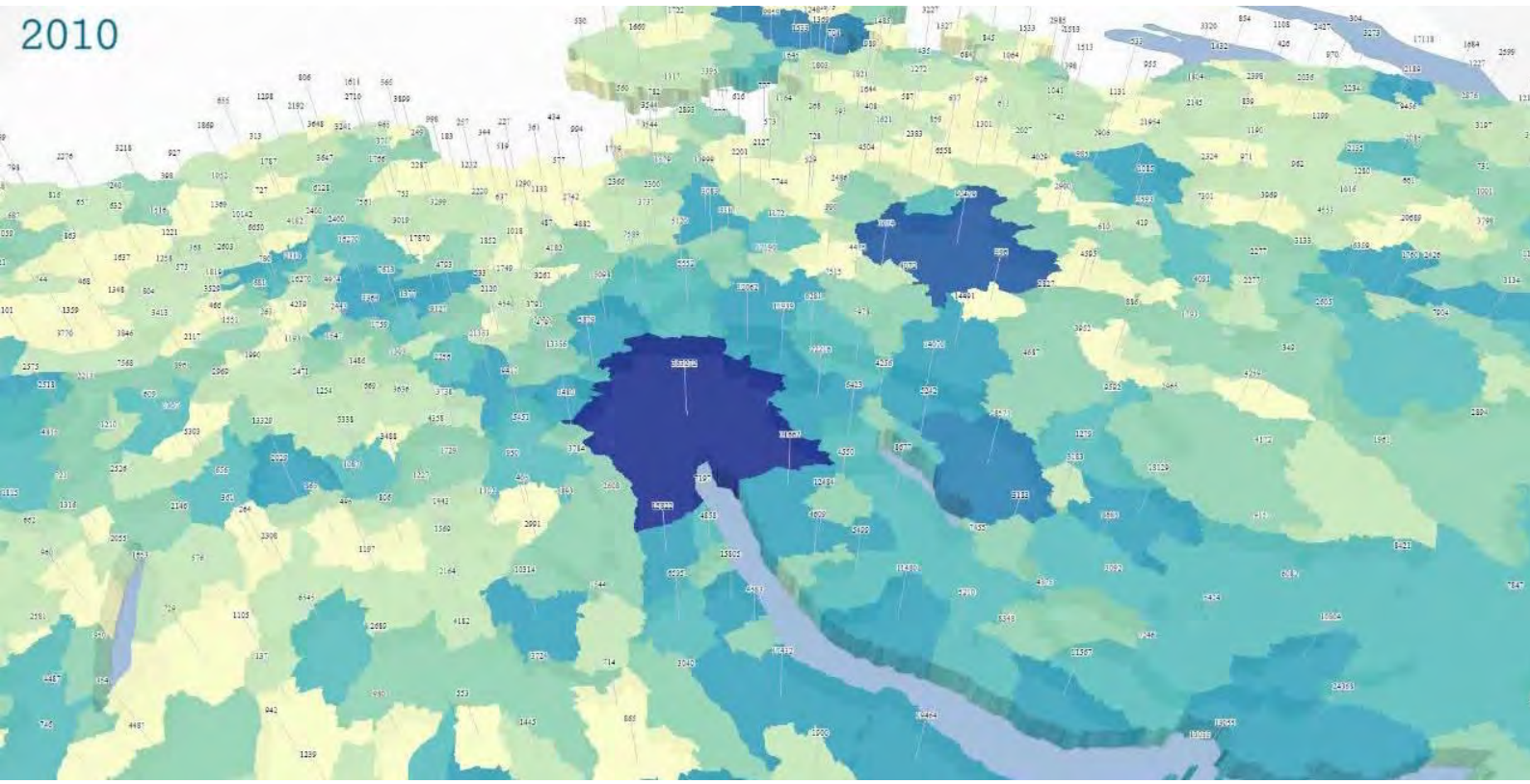
- Orte (Zonen)

Simulation Flächennutzung

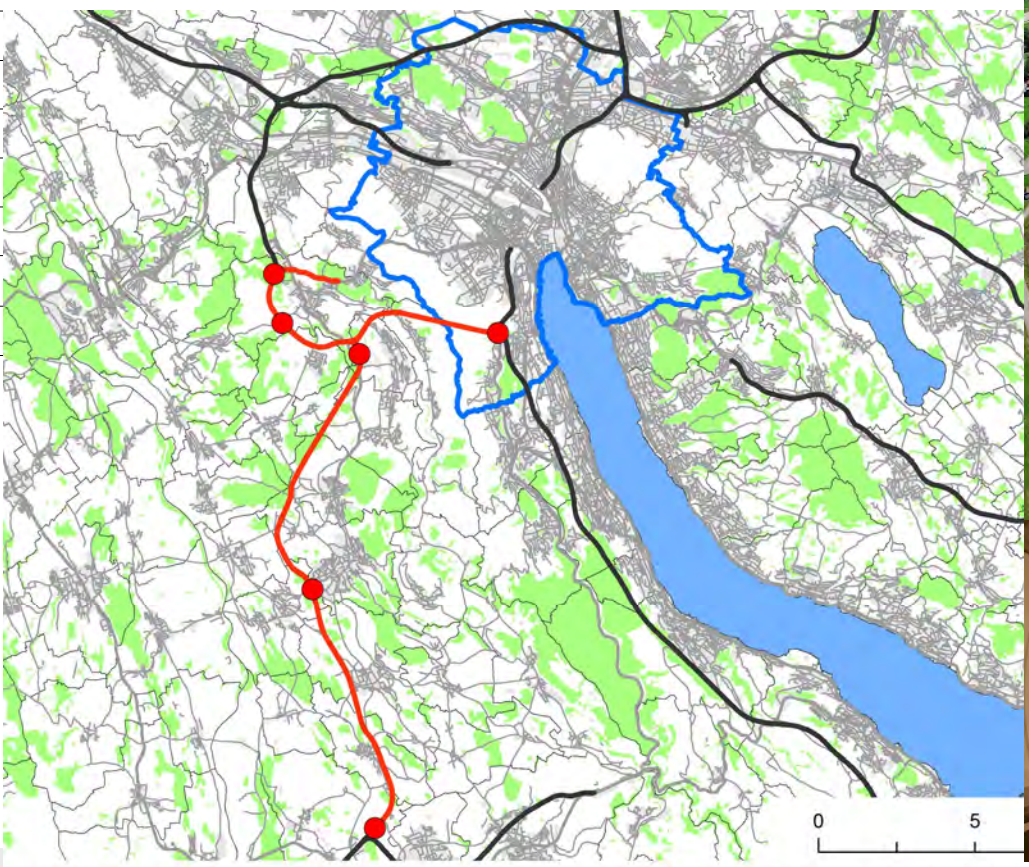
Jährliche Zyklen



Simulation Flächennutzung



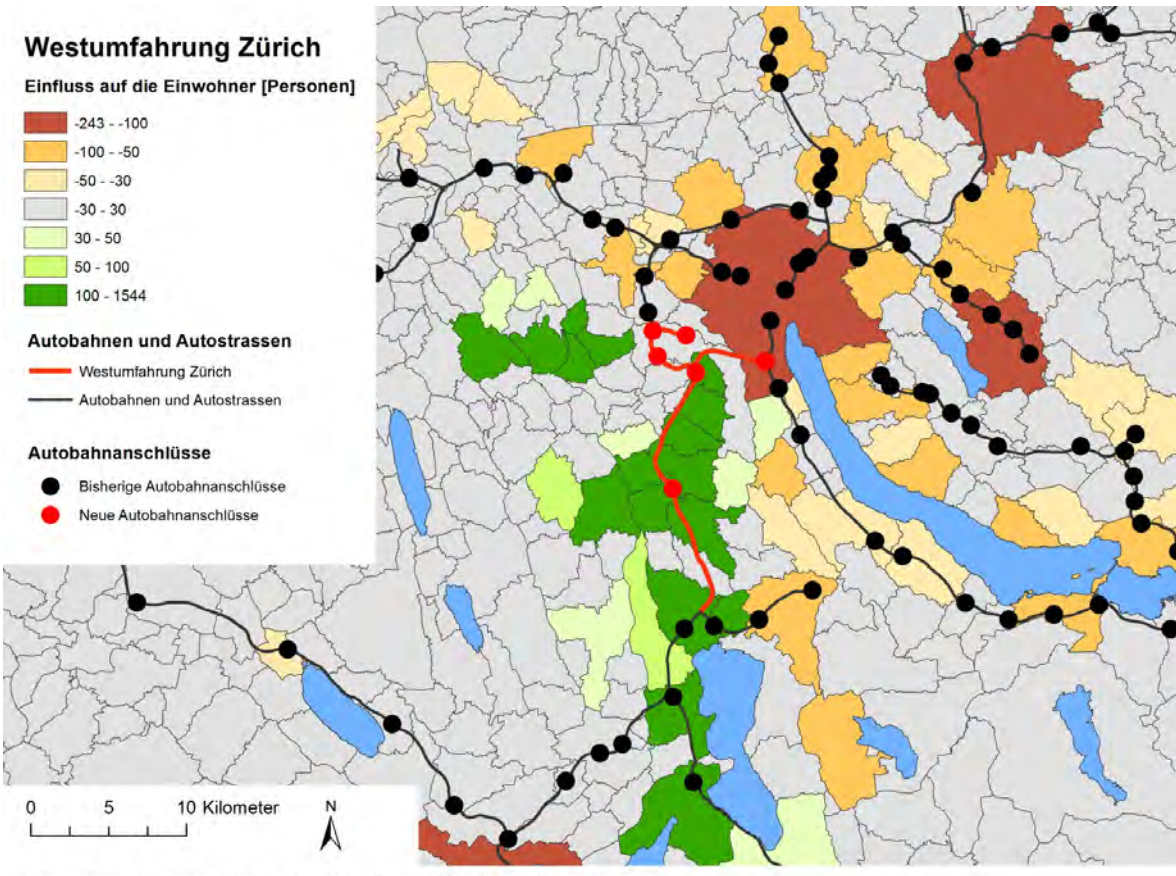
Westumfahrung Zürich



Quellen: Strassen © OpenStreetMap-Mitwirkende (OSMF, 2015); Gemeinden GG25 © Bundesamt für Landestopographie (2015); Siedlungsgebiet (Stand 2012) © Kantonale Raumplanungsfachstellen, INFOPLAN-ARE (2015)

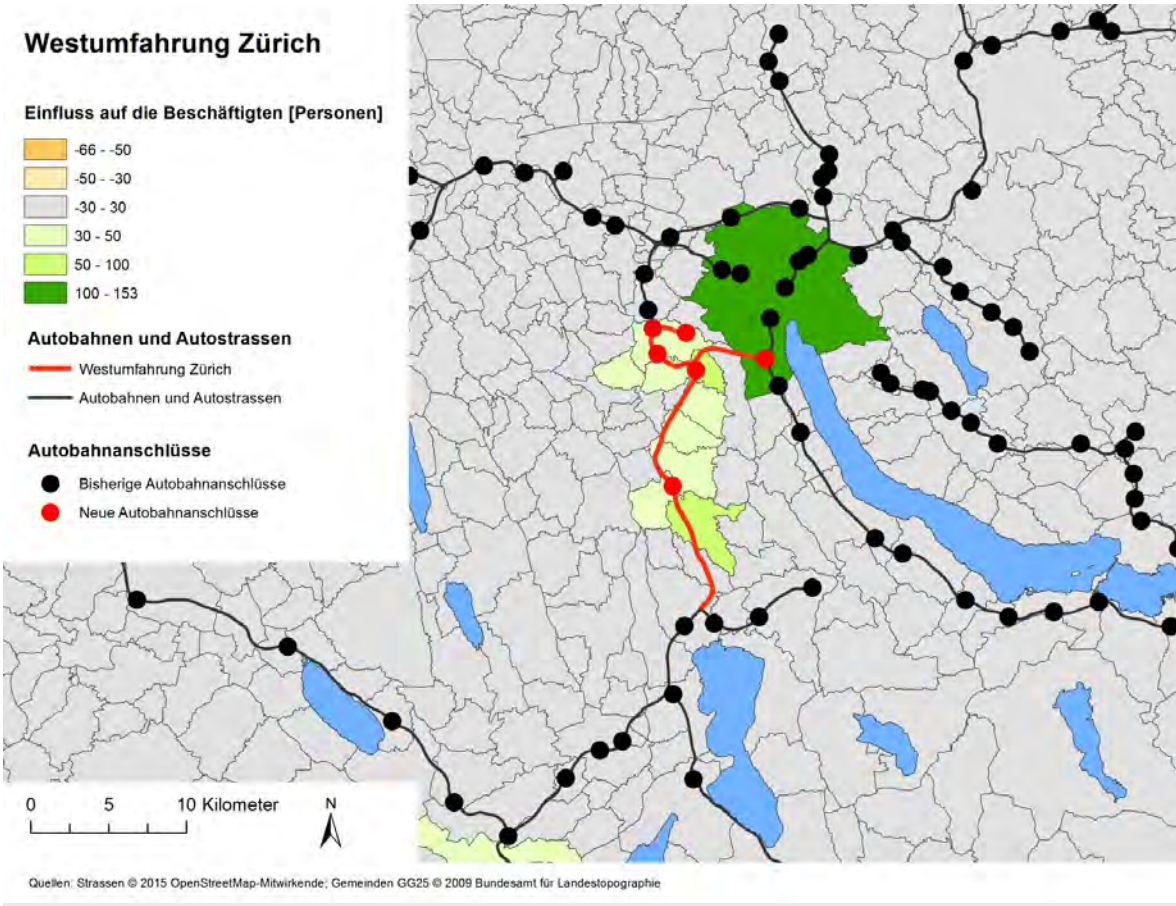
Wohnbevölkerung nach 10 Jahren

FaLC-sim



Quellen: Strassen © 2015 OpenStreetMap-Mitwirkende; Gemeinden GG25 © 2009 Bundesamt für Landestopographie

Beschäftigte nach 10 Jahren

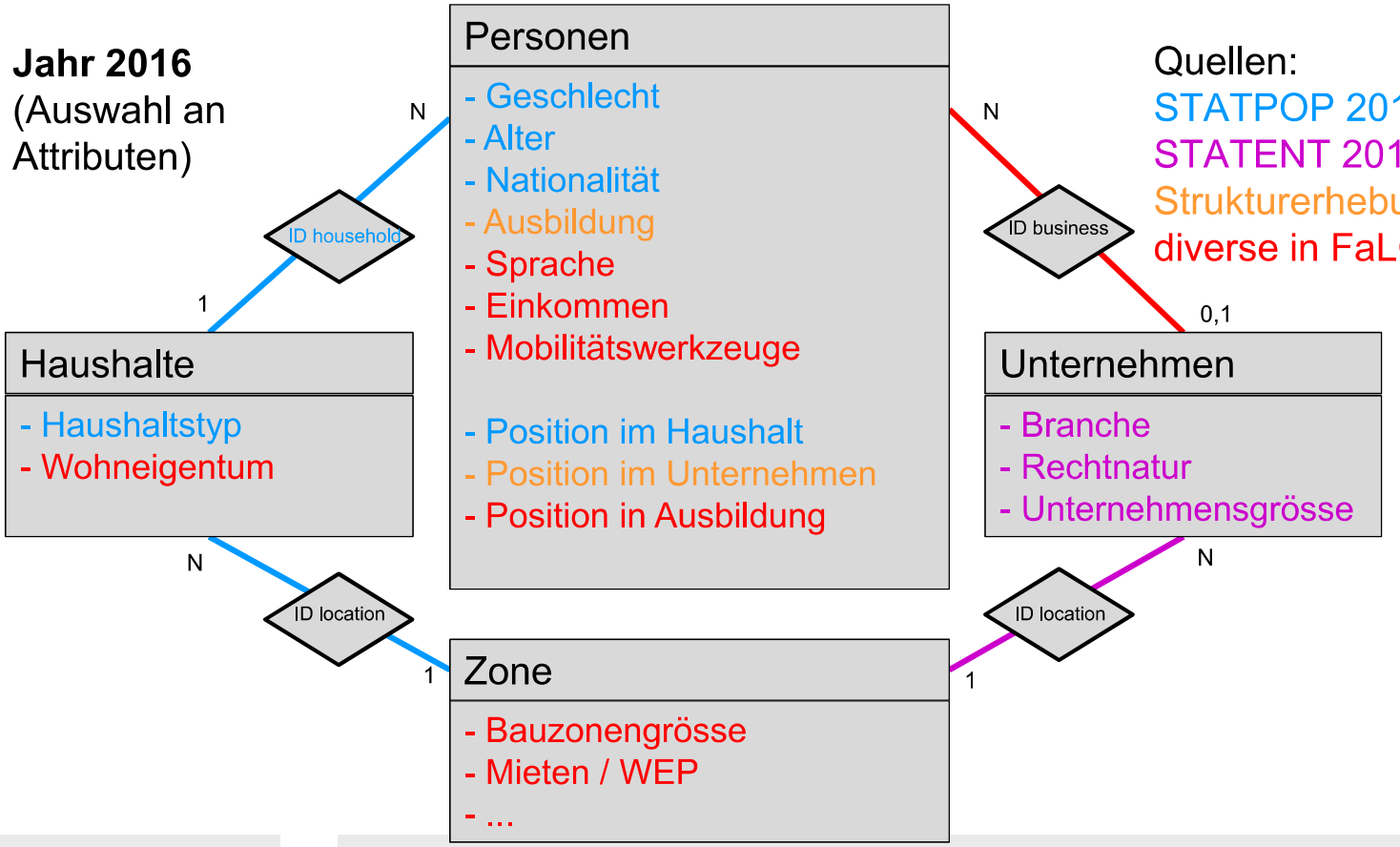


FaLC: Synthetische Population ...

... für die Schweiz .

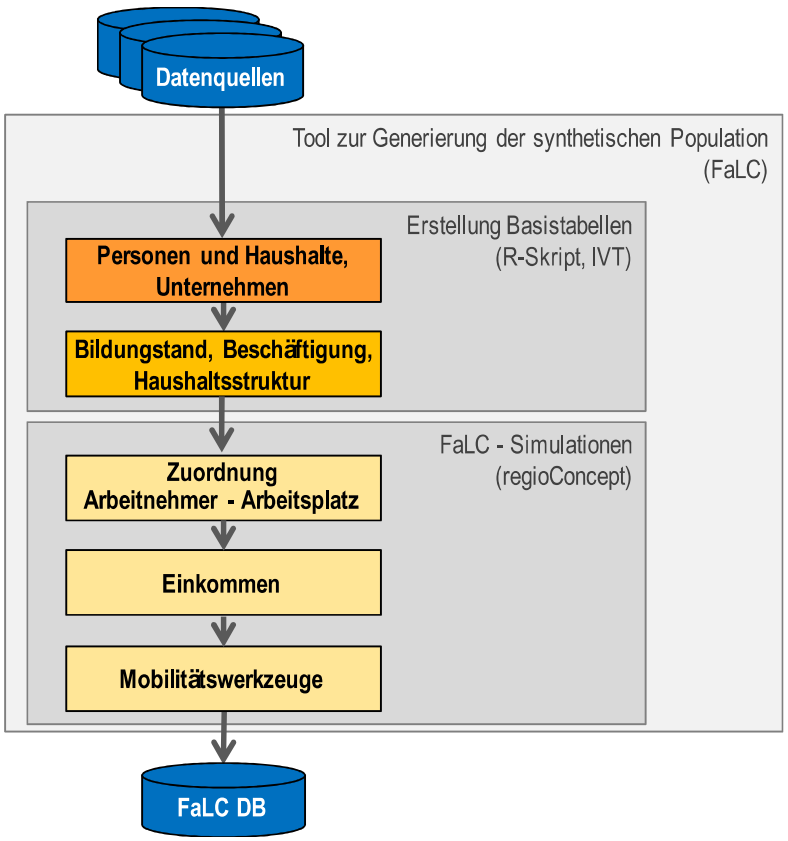
FaLC-Synpop für NPVM

Jahr 2016
(Auswahl an Attributen)



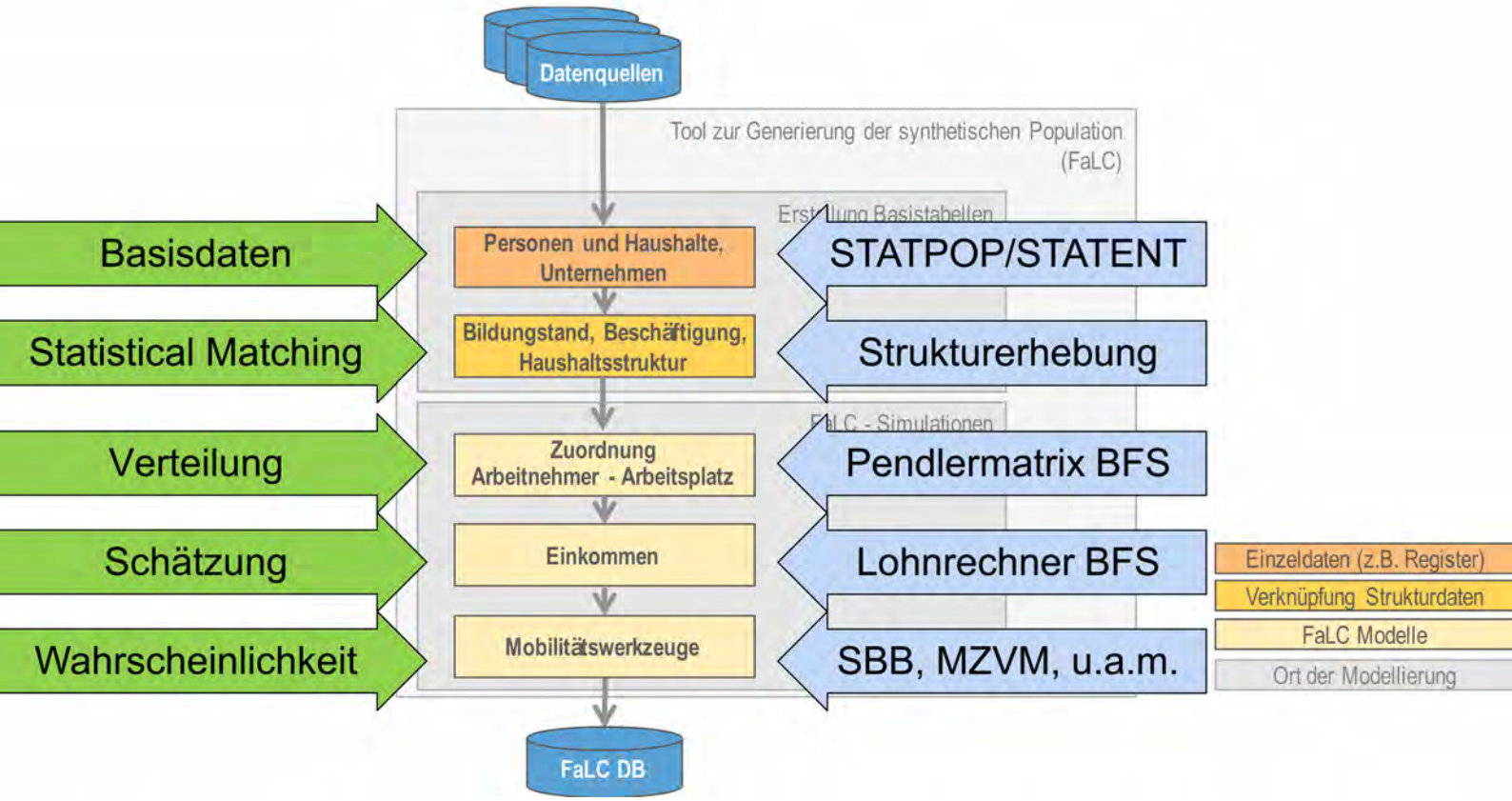
Quellen:
STATPOP 2016
STATENT 2015
Strukturerhebung
diverse in FaLC

Erstellung Synpop



Einzeldaten (z.B. Register)
Verknüpfung Strukturdaten
FaLC Modelle
Ort der Modellierung

Wichtigste Modellansätze



Typisches Beispiel…

… Mobilitätswerkzeuge .

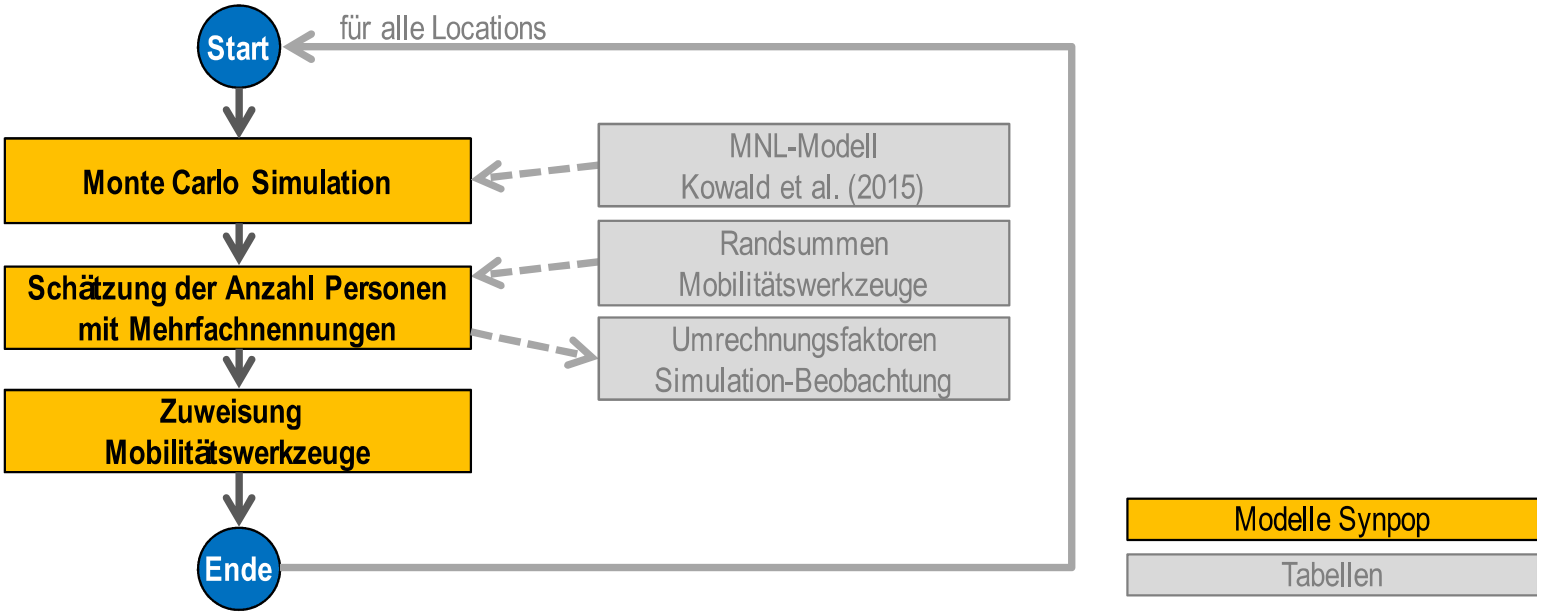
Code	Werkzeug	A priori-Annahmen	Anz. 2014* in Tausend	Priorität**
1	Auto + GA		400	1
2	Auto + Halbtax		1'984	7
3	Auto		2'876	9
20	Auto + Halbtax + Verbundsabo		101	3
30	Auto + Verbundsabo		51	4
4	GA		319	2
5	Halbtax		714	8
6	Kein Mobilitätswerkzeug		1'667	10
50	Halbtax + Verbundsabo		70	5
60	Verbundsabo		55	6

* Anz. 2014: Anzahl Personen in der Schweiz, Schätzung aufgrund der Verteilung gemäss Mikrozensus Mobilitätsverhalten MZMV 2010 (BFS, 2012)

** Priorität: Reihenfolge der Zuweisung

Typisches Beispiel…

… Mobilitätswerkzeuge .



Validierung

Information	CH	Reg.	Gem.	Zonen	Entität
Personen: Alter, Geschlecht	+++	+++	+++	+++	+++
Haushalte: Struktur	+++	+++	+++	+++	+++
Unternehmen (2000): Branche, Grösse	++	++	++	+	Zufall
Unternehmen (2014): Branche, Grösse	+++	+++	+++	+++	+++
Zuordnung Arbeitsplatz	++	++	+	+	Prob.
Bildungsstand, Beschäftigung	++	+	+	+	Prob.
Sprache (2000)	+++	+++	+++	+++	+++
Sprache (2014)	++	++	+	+	Zufall
Zuordnung Arbeitsplatz	++	++	+	+	Prob.
Einkommen	++	++	+/++	+/++	Prob.
Eigentum	++	++	+	+	Prob.
Mobilitätswerkzeuge (Besitz PW, Abos)	++	++	++	++	Prob.
Mobilitätswerkzeug (Verfügbarkeit PW)	+	+	+	+	Prob.

CH	Ganze Schweiz	+++	Identisch mit Grundgesamtheit
Reg.	Regionen/Bezirke/Kantone	++	Kleinste Abweichungen möglich (<1%)
Gem.	Gemeinden	+	Kleine Abweichungen möglich (<5%)
Zonen	NPVM-Zonen/Locations	Prob.	Zuordnung mit Wahrscheinlichkeiten
Entität	Person/Haushalt/Unternehmen	Zufall	Zuordnung zufällig

Fazit: Synthetische Population …

… nützt uns allen .

Pro und …

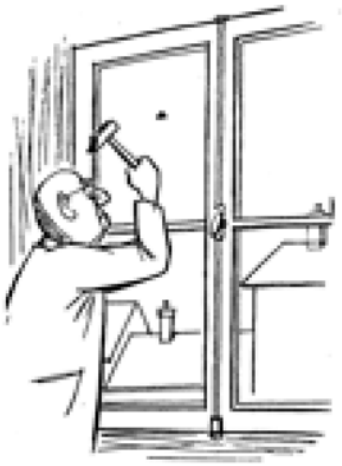
… Kontra .

Viele Chancen

- vereinfacht Informationsbeschaffung
  - Daten für Auswertungen / Planungsgrundlagen
  - Mehr Informationen für Planungen
- ermöglicht Vergleichbarkeit
  - gleiche Datengrundlage in verschiedenen Aufgaben
  - Standardisierung führt (meist) zu Transparenz
- gewährleistet höhere Qualität
  - Einbezug von mehr Informationen
  - Automatisierung

Pro und …

… Kontra .

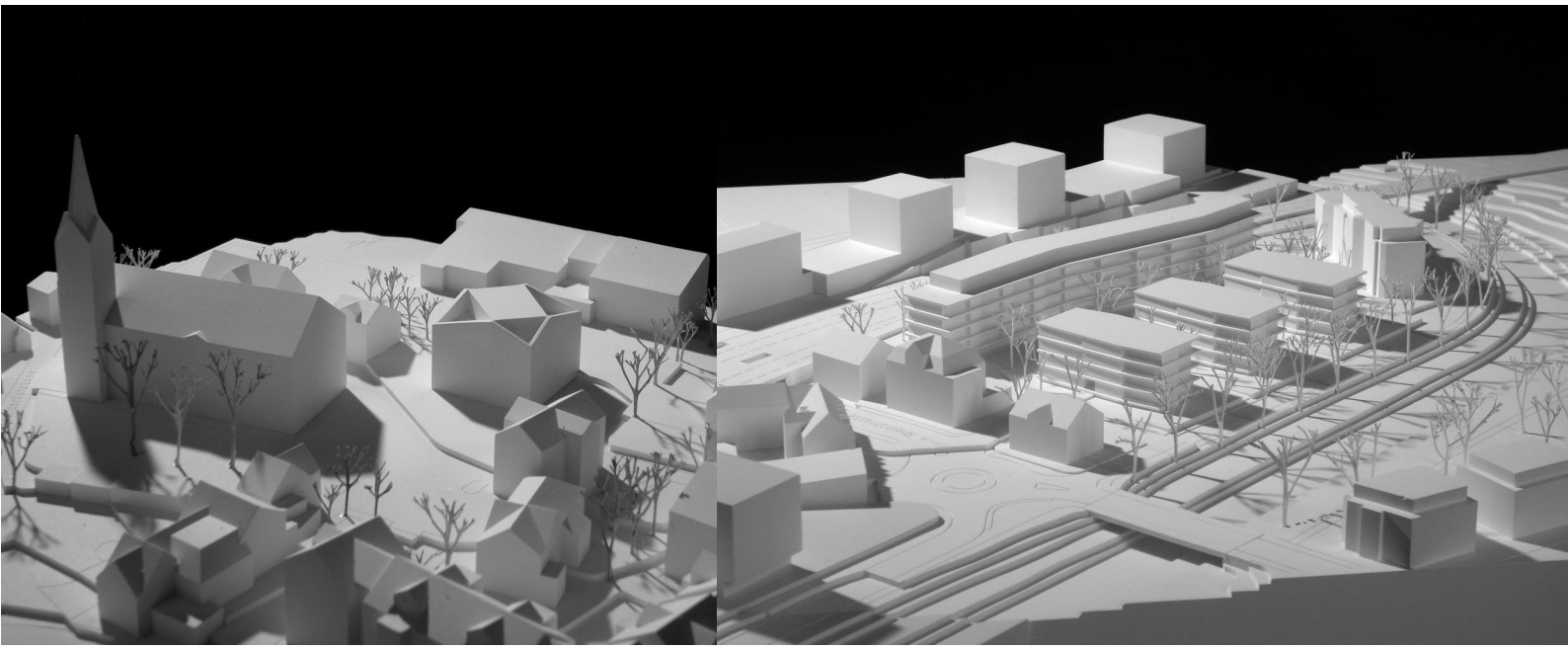


Einige Risiken

- Aufwand bei der Erst-Erstellung
  - hohe Kosten
- Datenverfügbarkeit
  - Wenn möglich werden Daten auf Personenebene genutzt
  - Datenschutz
  - konsistente Modellschätzungen
- Umgang mit Ungenauigkeiten
  - Zufall (weisses Rauschen)
  - Modellannahmen sind nicht für alle Fragen kohärent

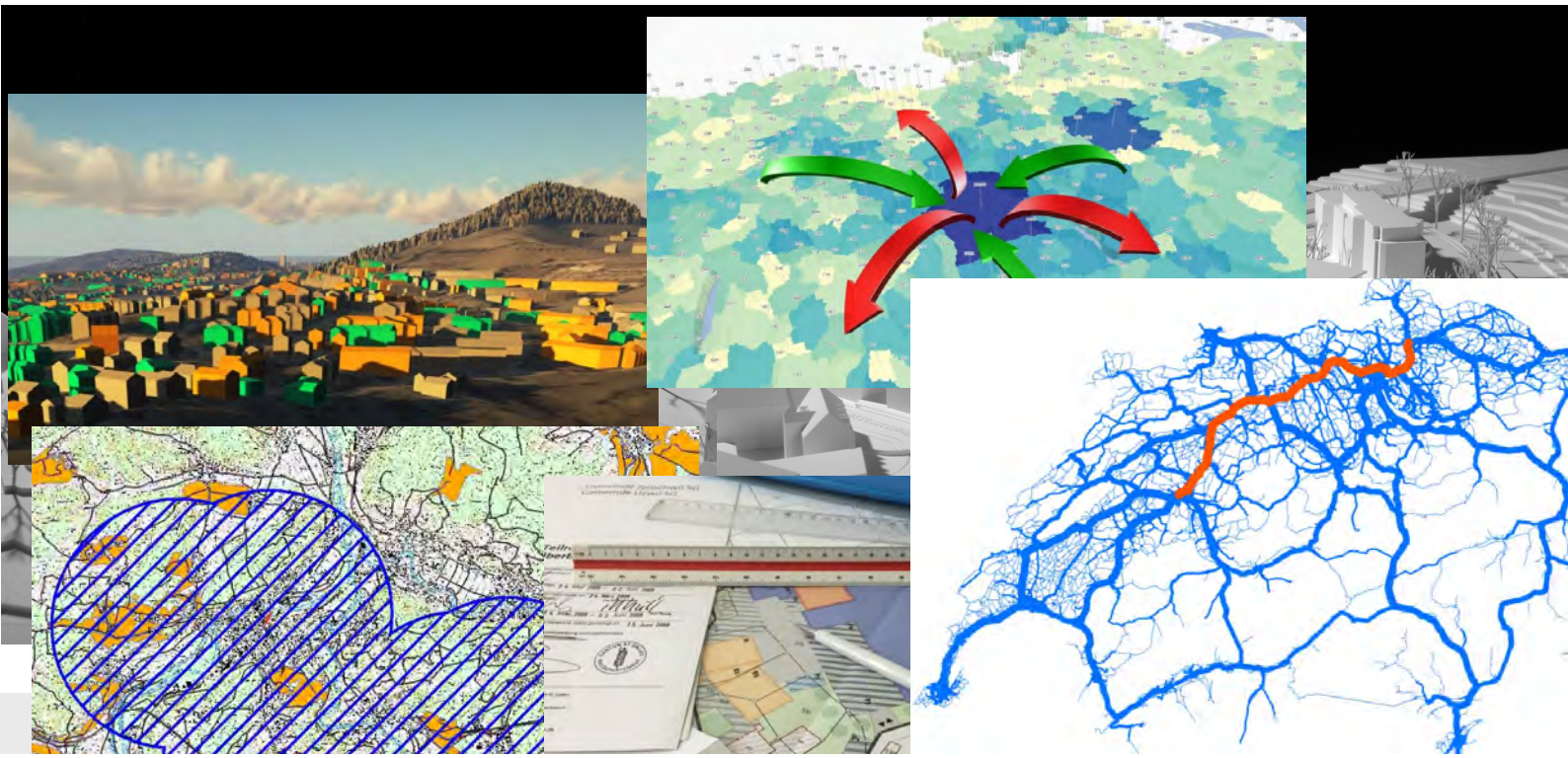
Fazit …

… Modell bleibt Modell .



Fazit …

… Modell bleibt Modell .



Zum Schluss …

Vielen Dank für die Aufmerksamkeit!

Kontakt:
Balz Bodenmann
balz.bodenmann@strittmatter-partner.ch

Weitere Informationen

Berichte zu FaLC

www.falc-sim.org

Projekte Strittmmatter Partner AG

www.strittmatter-partner.ch

Resultate Umfahrung Zürich

Bodenmann B.R. und P. Bürki (im Erscheinen) Räumliche Effekte der Mobilität auf die Verkehrs- und Siedlungsentwicklung, in M. Behnisch (Hrsg.) Flächenanspruchnahmen in Deutschland - beiträge zur quantitativen Methodik in der Raumwissenschaft, Springer, Heidelberg.

Wie schweizerisch ist die synthetische EBP-Schweiz?

Dr. Michel Müller, Silvan Rosser, Frank Bruns, Dr. Peter de Haan

Ziel des Inputs EBP

EBP: **synthetische Schweiz von EBP** vorstellen

Kunde: **Inputs für mögliche Best-Practice** in der Anwendung von synthetischen Populationen

Inhaltsverzeichnis

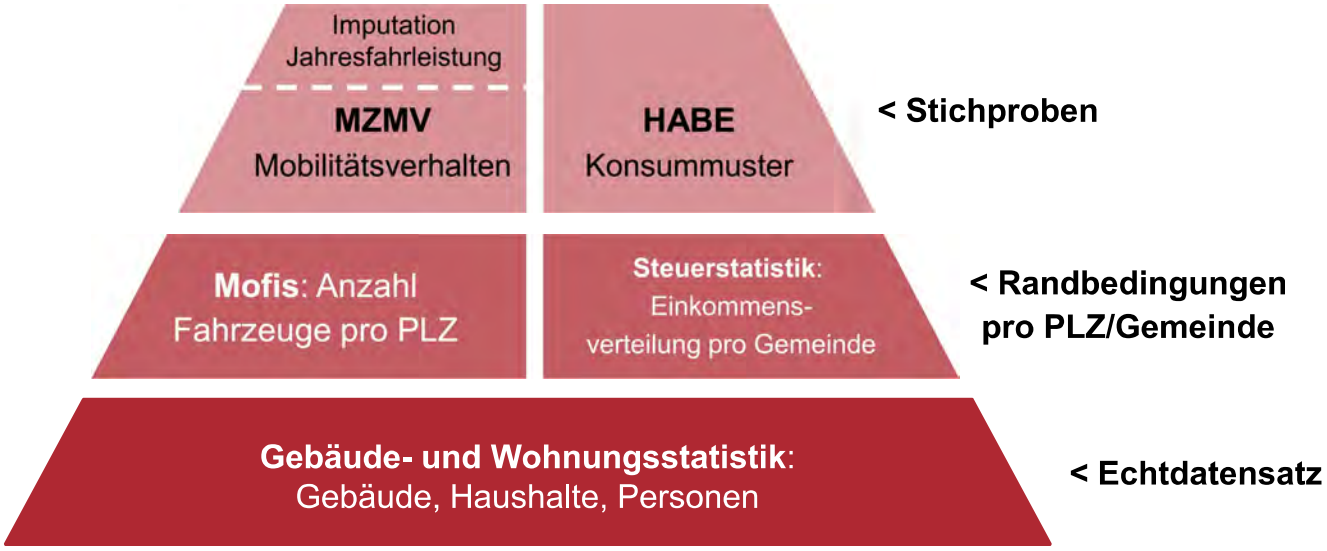
1. Die synthetische Schweiz von EBP ist…

Eine **«synthetische» schweizerische Bevölkerung**:

- **8 Millionen Personen** in ca. 3.6 Millionen Haushalten,
- mit einem **Wohngebäude**
- mit «**Mobilitätswerkzeugen**» (Autos und ÖV-Abonnemente) und zugehöriger Mobilität (Personenkilometer MIV und ÖV)
- mit einer **Einkommen**sklasse sowie **Konsum**gewohnheiten (Ausgaben pro Konsumkategorie)

➔ Möglichst intelligente Verknüpfung von Wohnen, Mobilität, Konsum

1. Die synthetische Schweiz von EBP beruht auf…



1. Synthetische Schweiz von EBP, Aufbau…

Das Rückgrat: GWS, pro Hektare

- Wohnungen je Gebäude und Anzahl Personen je Haushalt, mit vorhandenen Alterskategorien der Haushaltangehörigen (bis 18, 18–25, 25–65, >65)
- Grundtabellen: Hektaren, Gemeinden, Gebäude, Wohnungen, Personengruppe [Personenhaushalte], Personen
- GWS, **imputiert**: genaue Anzahl Personen je Alterskategorie
- GWS, **imputiert**: 7 HH-Typen (siehe nächste Folie)

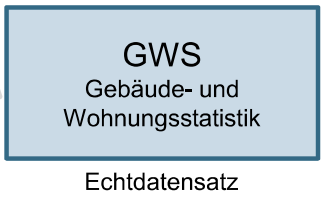
GWS
Gebäude- und
Wohnungsstatistik

Echtdatensatz

1. Synthetische Schweiz von EBP, Aufbau…

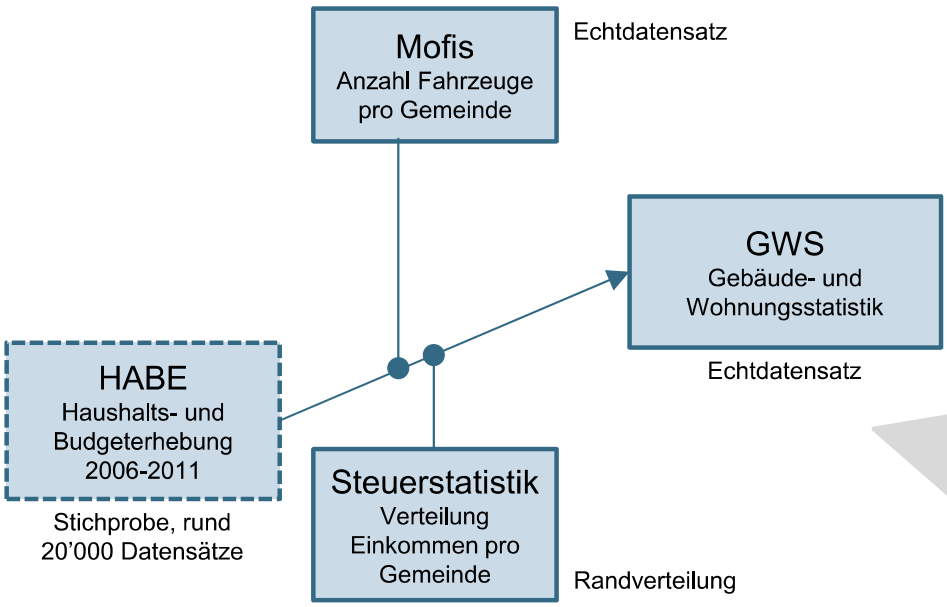
7 Haushaltstypen:

- *Cou*: Paare, unter 65
- *Fam_erwachs*: Familien mit Jungerwachsenen 18-24, keine Minderjährige, inkl. Alleinerziehende
- *Fam_kinder*: Familien mit minderjährigen Kindern, inkl. Alleinerziehende
- *MPH*: Mehrpersonenhaushalte
- *Sen_cou*: Seniorenpaare, beide > 65
- *Sen_sin*: Senioren Einpersonenhaushalt, > 65
- *Sin*: Einpersonenhaushalt, < 65



Echtdatensatz

1. Synthetische Schweiz von EBP, Aufbau…

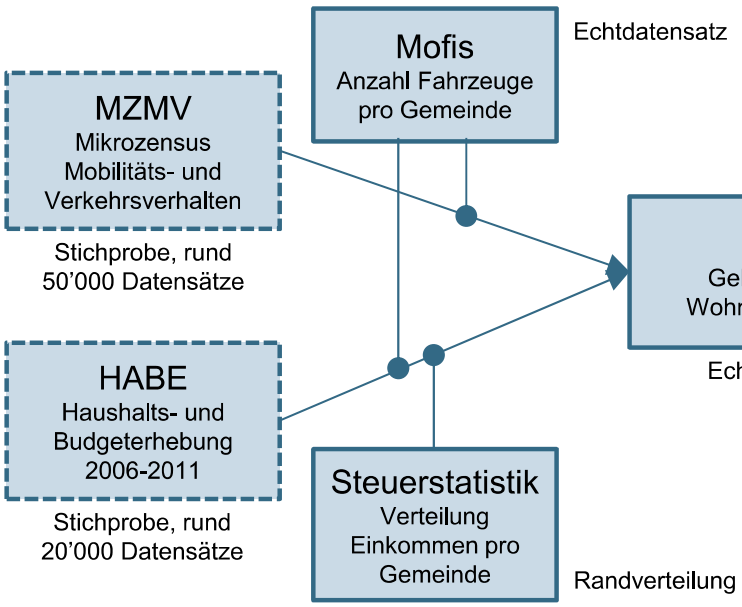


HABE Zuordnungsmodell

Matching an die ~3.6 Mio. Haushalte, Indikatoren:

- Haushaltstyp
- Autohaushalt (0/1+)
- Gemeindetyp (Stadt/Agglo/ländlich)
- Verfügbares Einkommen (4 Einkommensklassen)
- BFS-Grossregionen

1. Synthetische Schweiz von EBP, Aufbau…

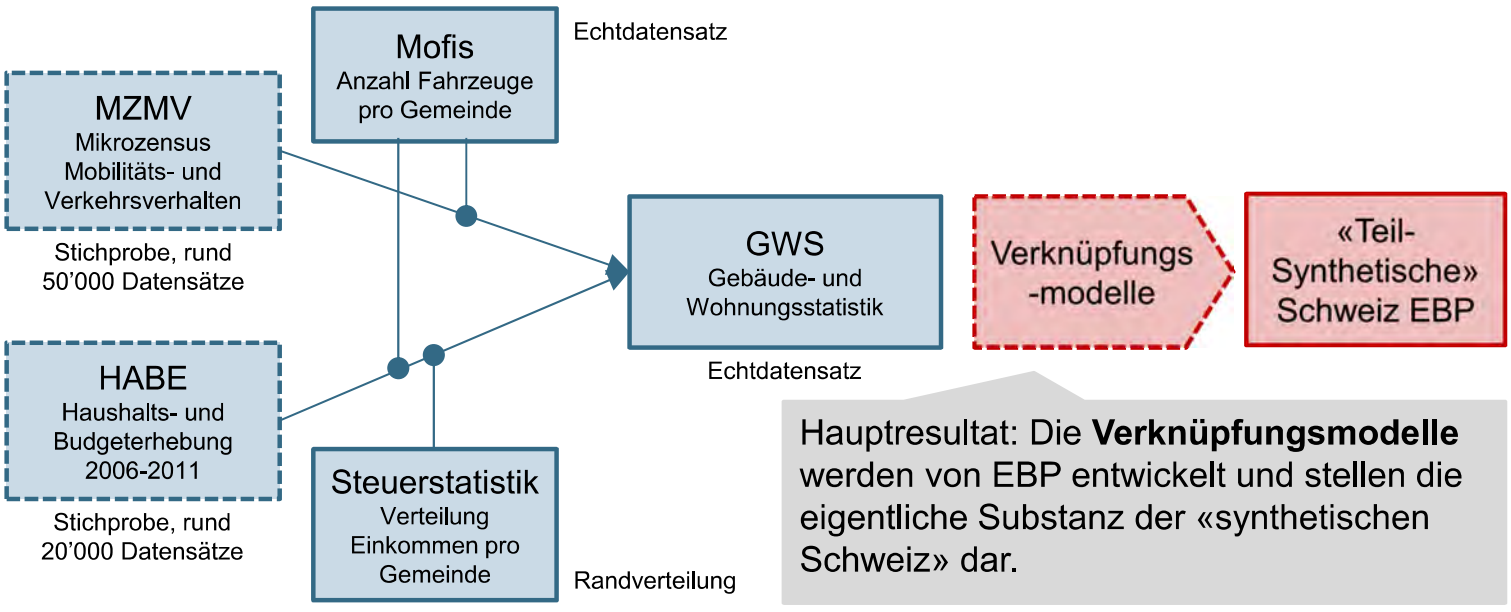


MZMV Zuordnungsmodell

Matching an die ~3.6 Mio. Haushalte, Indikatoren:

- GWS: Haushaltstyp
- MOFIS > GWS/HABE: 0/1/2+ Autos
- Gemeindetyp (Stadt/Agglo/ländlich)
- BFS-Grossregionen

1. Synthetische Schweiz von EBP, Aufbau…



Hauptresultat: Die **Verknüpfungsmodelle** werden von EBP entwickelt und stellen die eigentliche Substanz der «synthetischen Schweiz» dar.

2. Chancen

Die synthetische Schweiz bietet Mehrwert in vielfältigen Projekten von EBP:

- **Komplettierung statistischer Grundlagendaten**: Anhand von bestehenden Umfragen Lücken füllen. Insbesondere für geographische Räume mit zu geringen Stichproben
- **Evaluationen von Politikmassnahmen**: Wirkungen von Massnahmen können nach geographischen und sozio-demographischen Strukturen beschrieben werden ➔ Verteilungswirkungen
- **Bessere Entscheidungsgrundlagen**: Verbesserung von Entscheidungen mit räumlich struktureller Komponente (bspw. Planung Mobilitätsangebote)
- **Kommunikationskampagnen**: gezieltere und effizientere Ansprache von Kundengruppen nach geographischer Struktur

3. Risiken

Das Produkt «synthetische Population» ist anspruchsvoll. Neben den Chancen muss mit Risiken explizit umgegangen werden:

- **Lost in Data**: zusätzliche Daten bringen nicht per se einen Mehrwert. Ohne klare Definition von Umgang und Anwendung der Daten kann eine detaillierte Grundlage wie die synthetische Population auch kontraproduktiv Verwirrung stiften
- **Nutzen berechtigt den höheren Aufwand nicht**: Die Erstellung einer synthetischen Population ist aufwändig, der Nutzen muss diesen Aufwand rechtfertigen
- **Inkonsistenzen zu anderen Datenquellen**: Die synthetische Population ist eine neue Datenquelle. Sie kann in Konkurrenz und Widerspruch zu anderen Datenquellen stehen.

EBP

These Anwendung in der Schweiz

In der Schweiz besteht eine grosse Chance, teilsynthetische Populationen anzuwenden

– Hohe Güte der Grundlagendaten (Register- und Strukturerhebung) ermöglicht die (teilweise) Verwendung von Echtdaten

Für die Anwendung in der Praxis aufzeigen:

– Best Practice für Methodik von teilsynthetischen Populationen
– Best Practice für Anwendungen, wo liegen die grössten Potenziale
– klar aufzeigen, welche Eckpunkte für die Verwendung bestehen: Verfügbarkeit von Daten, Lizenzen, Datenschutz, etc.

EBP

These zum Besteller—Lieferanten Verhältnis

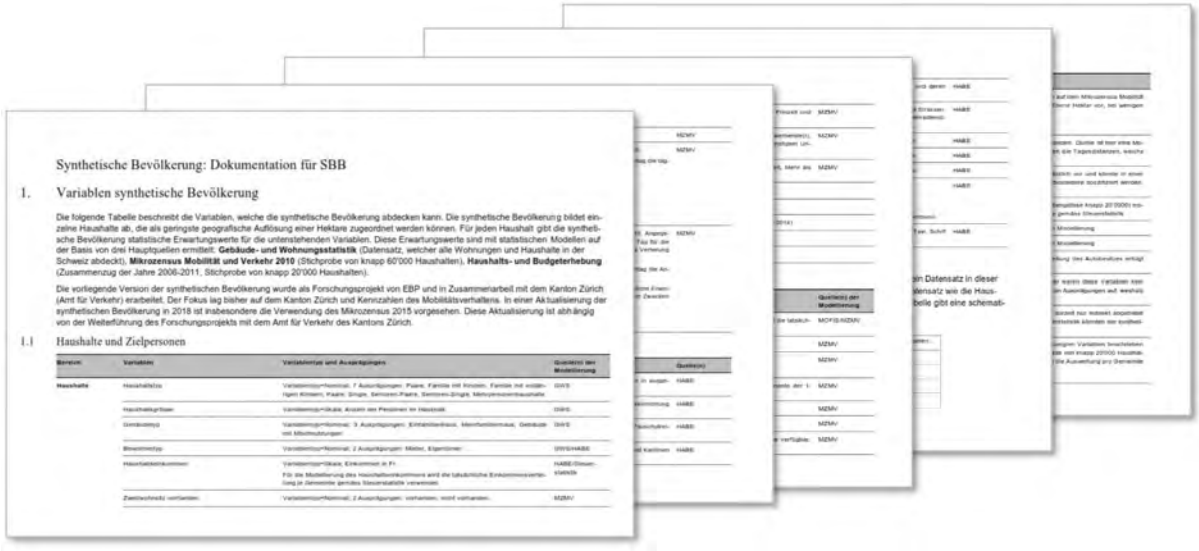
Die Bestellung einer synthetischen Population bedarf einer detaillierten vorgängigen Einigung zu Anwendung, Inhalt und Lieferung

Kernpunkte für diese Einigung umfassen idealerweise:

– beabsichtigte **Anwendung**
– verwendete **Methoden**
– zu liefernde **Variablen**, darin enthaltene **Ausprägungen**
– **Format** der Lieferung, Verwendbarkeit beim Besteller

These zum Besteller—Lieferanten Verhältnis

Beispiel einer solchen Dokumentation (Besteller: SBB, Lieferant: EBP):


Synthetische Bevölkerung: Dokumentation für SBB

1. Variablen synthetische Bevölkerung

1.1 Haushalte und Zielpersonen

These Qualitätssicherung

Vor Lieferung einer synthetischen Population sollten die Güte und die Methoden der Qualitätssicherung gemeinsam definiert werden

«**wie schweizerisch muss die synthetische Schweiz sein?**»

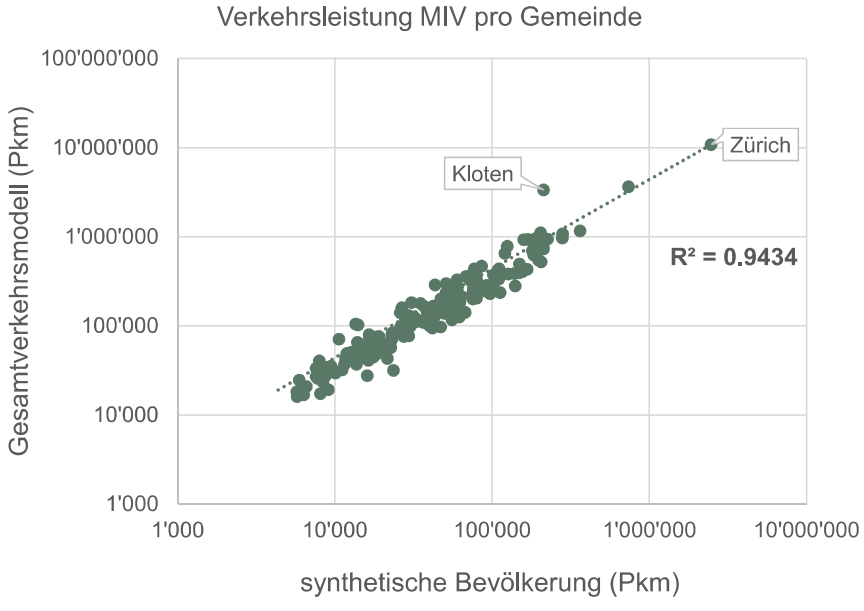
Festgelegt werden sollte:

- Die Güte der Daten: Kriterien und Strukturen festhalten, was auf welcher Ebene wie gut stimmen soll.
- Die Methoden (und Zuständigkeiten) der Qualitätssicherung sollten vorgängig festgelegt werden
- Geeignetes Werkzeug zur Qualitätssicherung: Hypothesen
- Geeignetes Werkzeug zur Qualitätssicherung: Visualisierungen

Hypothesen zur Qualitätssicherung

- Idealerweise können zur Qualitätssicherung identische Indikatoren aus a) der synthetischen Population und b) einer unabhängigen zweiten Datenquelle verglichen werden
- Trotz «gleichem» Indikator bilden die Datenquellen nicht immer die gleichen Zusammenhänge ab
- Illustration an Beispiel des Vergleichs der Fahrleistung MIV im Kanton Zürich:
  a) synthetische Schweiz EBP, und
  b) Gesamtverkehrsmodell Kanton Zürich

Hypothesen zur Qualitätssicherung



Im **Vergleich nach Gemeinden** wird die Struktur der Verkehrsleistung gut abgebildet (statistisches Bestimmtheitsmass R^2 = 94%)

Die Steigung des Zusammenhangs in der Grafik und der Vergleich der gesamten kantonalen Fahrleistung zeigt jedoch, dass das Gesamtverkehrsmodell viel mehr MIV Mobilität enthält.

Für diese Abweichung wurden Hypothesen formuliert, mit welchen die Abweichung erklärt werden konnte

EBP

Hypothesen zur Qualitätssicherung

Verkehrsleistung MIV im Kanton Zürich			
GVM	54.686 Mio. km	Synthetische Bevölkerung	13.654 Mio. km

Auf den Vergleich oben müssen drei Korrekturen angewendet werden:

i) Bereinigung von Doppelzählungen im GVM,

ii) Bereinigung der synthetischen Bevölkerung aufgrund der Anzahl Zielpersonen, und

iii) Bereinigung aufgrund der unterschiedlichen Mobilität, die im GVM bzw. dem Mikrozensus (als Grundlage der synthetischen Bevölkerung) abgebildet ist.

Verkehrsleistung MIV im Kanton Zürich (korrigierter Vergleich)			
GVM	**38.067 Mio. km**	Synthetische Bevölkerung	**38.997 Mio. km** 2.4% Abweichung

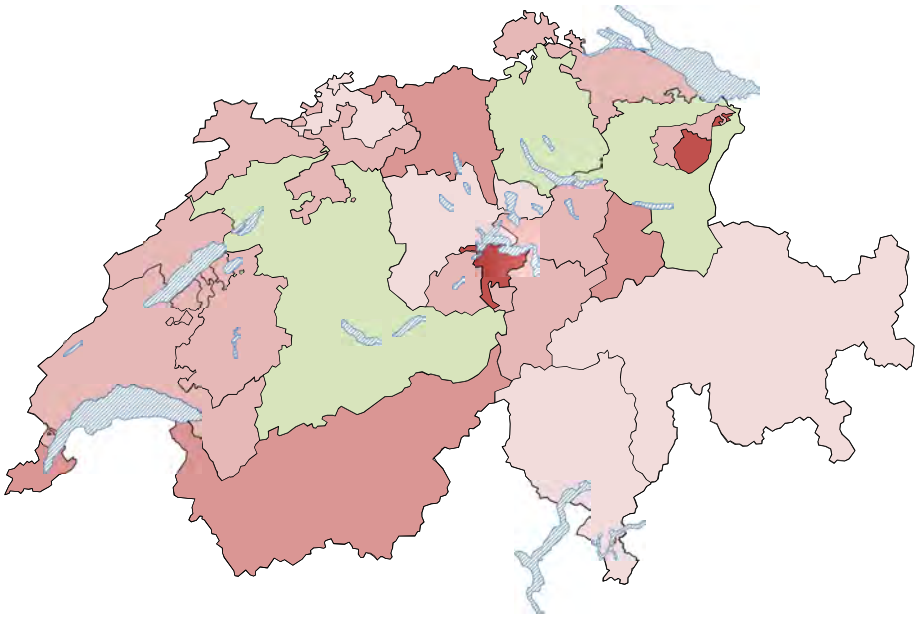
EBP

Visualisierungen zur Qualitätssicherung

- eine erste schnelle Einschätzung der Güte kann mittels Visualisierungen erfolgen
- in Verbindung mit ausgewählten Kenngrössen geeignet zur Qualitätssicherung von synthetischen Populationen

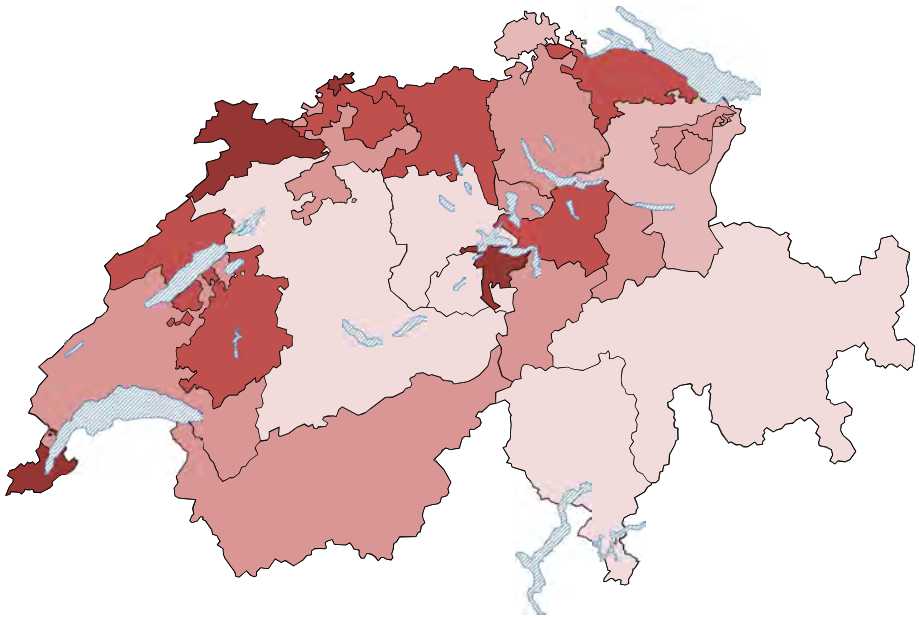
- Beispiel:
  - für den Indikator «mittlere Tagesdistanz MIV»
  - Auswertung nach Kantonen
  - Differenz der synthetischen Bevölkerung zum Mikrozensus vs.
  - Differenz eines Zufallsmodells zum Mikrozensus

Visualisierungen zur Qualitätssicherung



synCH / MZMV

Visualisierungen zur Qualitätssicherung



Zufall / MZMV

Vielen Dank für Ihre Aufmerksamkeit!

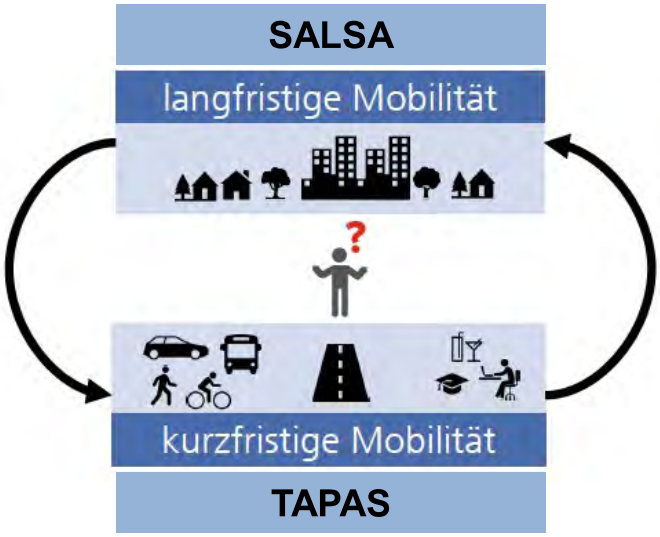
Übersicht der Thesen:

- In der Schweiz besteht eine grosse Chance, teilsynthetische Populationen anzuwenden
- Die Bestellung einer synthetischen Population bedarf einer detaillierten vorgängigen Einigung zu Anwendung, Inhalt und Lieferung
- Vor Lieferung einer synthetischen Population sollten die Güte und die Methoden der Qualitätssicherung gemeinsam definiert werden: «wie schweizerisch muss die synthetische Schweiz sein?»

Generierung synthetischer Bevölkerungen für Berlin - Möglichkeiten und Grenzen

Rita Cyganski, Antje von Schmidt, Benjamin Heldt
DLR-Institut für Verkehrsforschung

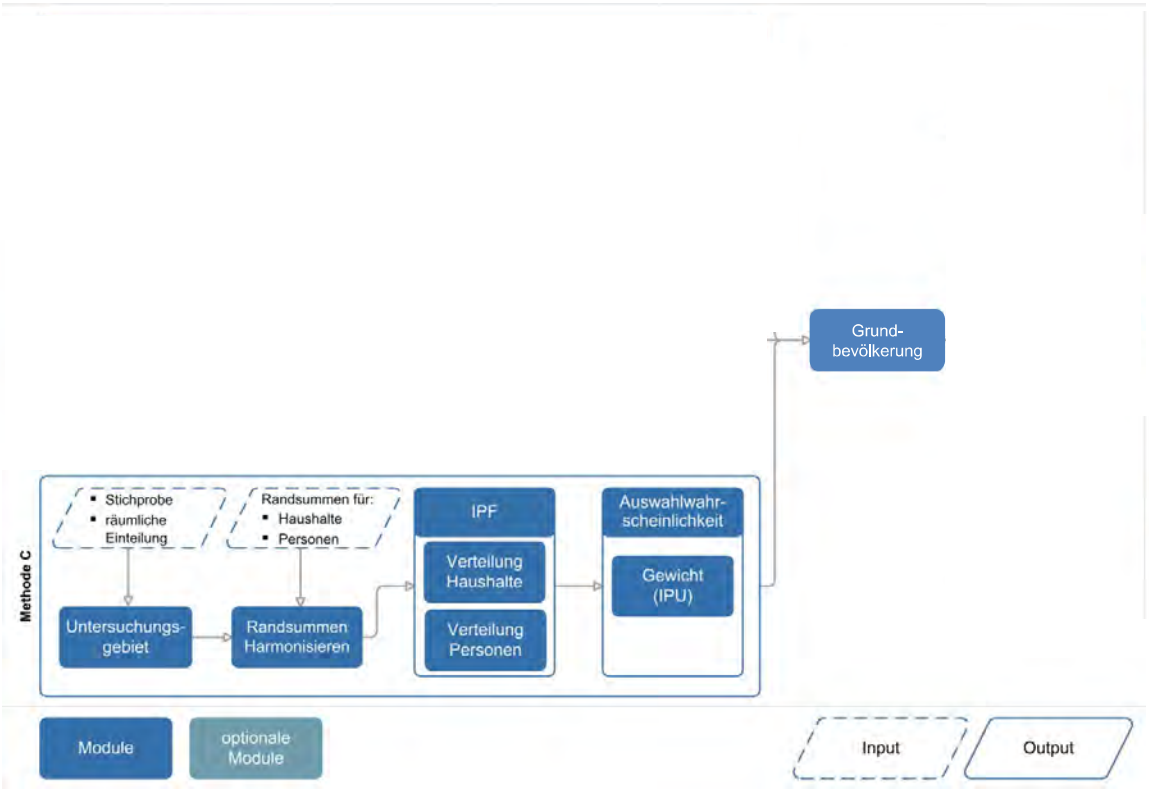
Synthetische Bevölkerung – warum die Mühe?



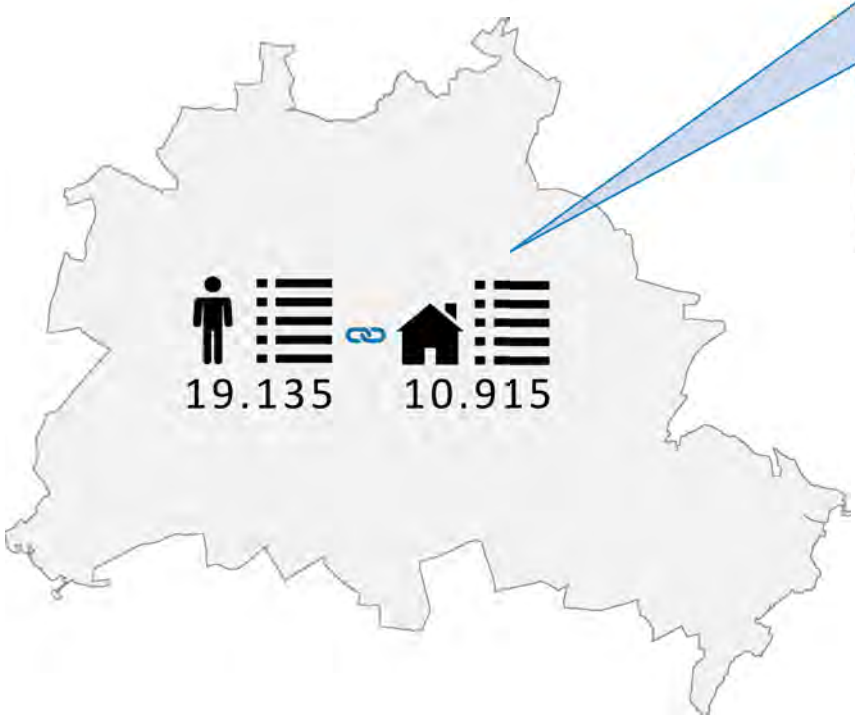
Perspektivische Nutzung

- Möglichkeit der aggregierten Nutzung in makroskopischen Modellen (PTV VISUM)
- Unternehmensstandortwahl (SALSA)
- Fortschreibung von Strukturgrößen, z.B. Standorte des Lebensmittelhandels nach Größe und Angebotstyp

Erstellung synthetischer Bevölkerungen mit dem SYNTHESIZER



Unsere Ausgangsbasis: Stichprobendaten



Stichprobe
disaggregierte soziodemographische Haushalts- und Personendaten (**Berlin gesamt**) Quelle: **Mikrozensus 2010**

HH-ID	HH-Größe	Einkommen	...
10001	2 Personen	5000 €	...
...			

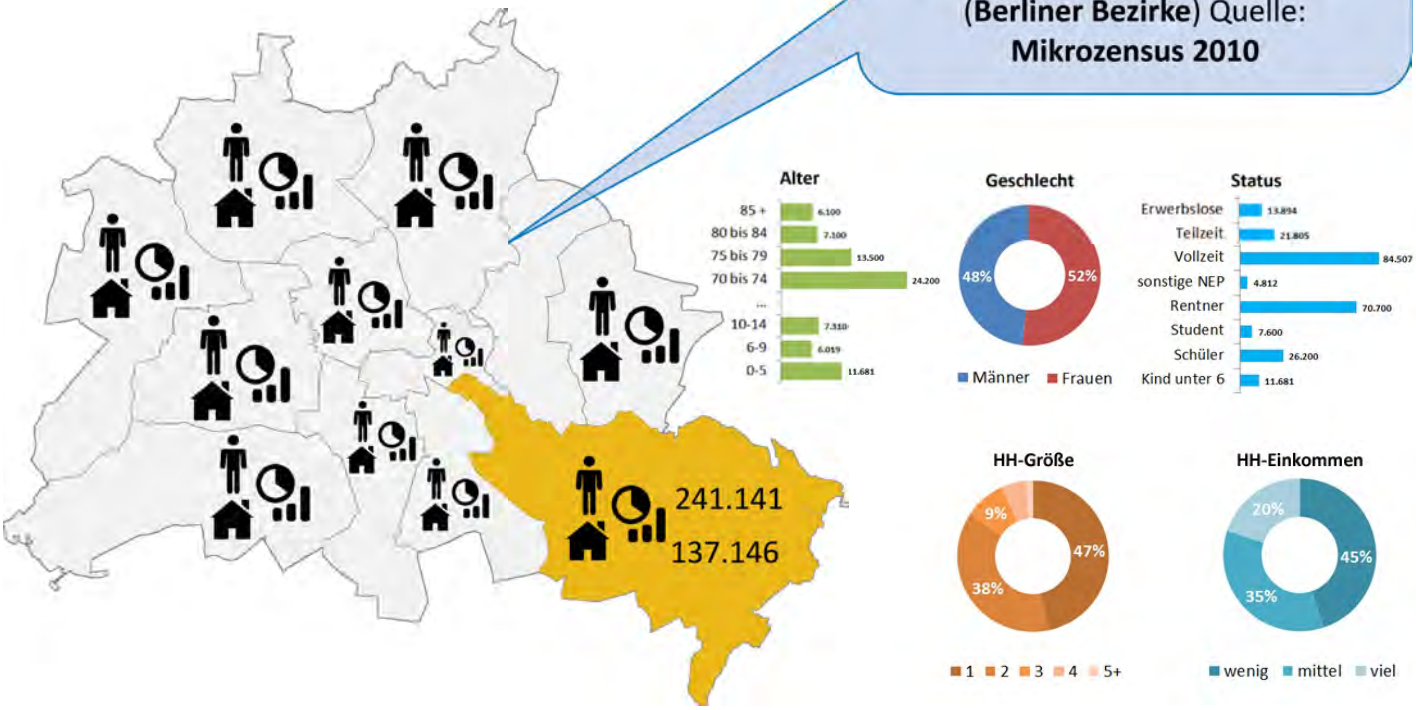
Haushalte

P-ID	HH-ID	Geschlecht	Alter	...
1000101	10001	männlich	52	...
1000102	10001	weiblich	48	...
...				

Personen

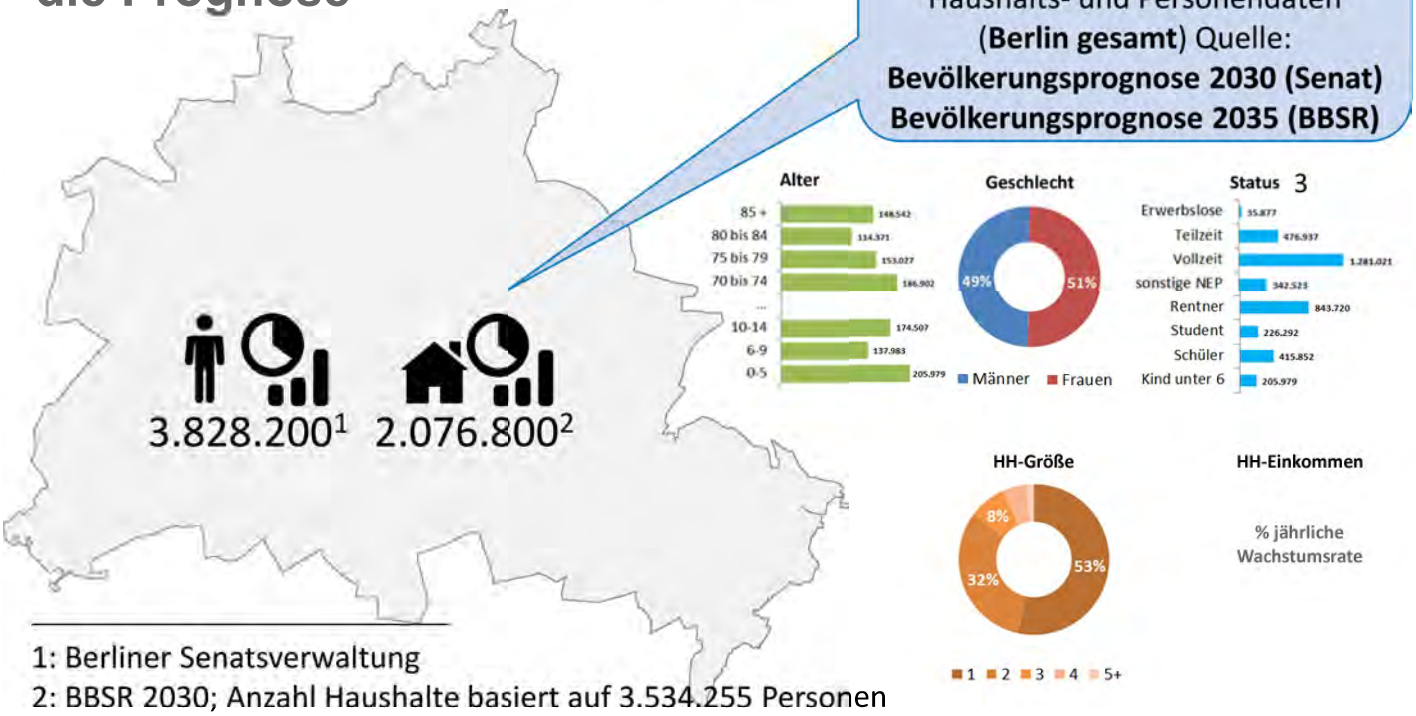
Festlegung der Zielverteilung für die Referenz

Randsummen (Referenzjahr)
aggregierte soziodemographische Haushalts- und Personendaten (**Berliner Bezirke**) Quelle: **Mikrozensus 2010**



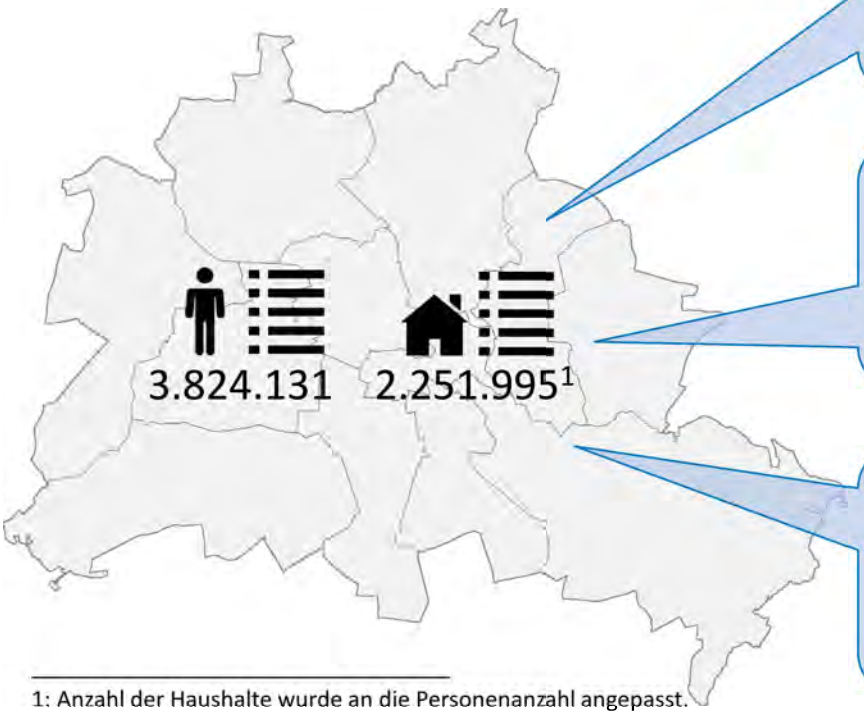
Festlegung der Zielverteilung für die Prognose

Randsummen (Prognosejahr)
aggregierte soziodemographische Haushalts- und Personendaten (**Berlin gesamt**) Quelle: **Bevölkerungsprognose 2030 (Senat)** **Bevölkerungsprognose 2035 (BBSR)**



1: Berliner Senatsverwaltung
2: BBSR 2030; Anzahl Haushalte basiert auf 3.534.255 Personen
3: Kinder, Schüler und Rentner gemäß Alter; rel. Änderung Stud.zahl (Kultusminister Konferenz); Erwerbspersonen (VEU2); 98% Beschäftigung (VP2030); Anteil Vollzeit/Teilzeit (Mikrozensus 2010)

Erweiterung der Grundbevölkerung



Grundbevölkerung
beinhaltet disaggregierte soziodemographische Daten

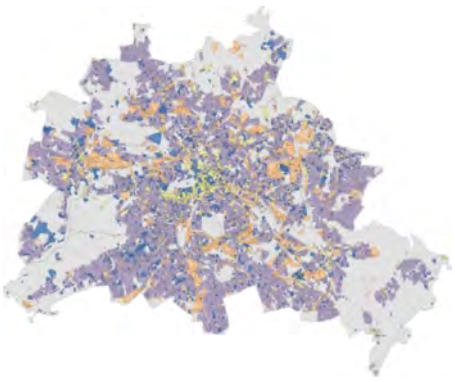
Räumliche Auflösung
basiert auf den verwendeten Randsummen (**Bezirksebene**) deshalb **Haushalte auf Adresskoordinaten verteilen**

Mobilitätsoptionen hinzufügen

Quelle:
Verkehrserhebungen[2]

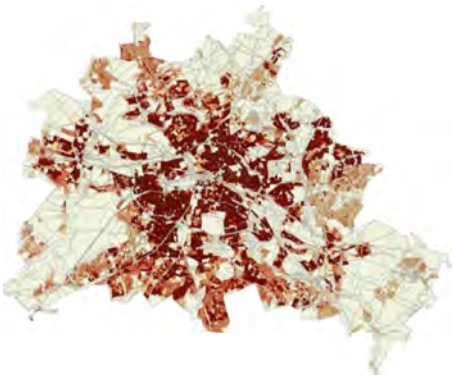
1: Anzahl der Haushalte wurde an die Personenanzahl angepasst.
2: Mobilität in Deutschland (MID), Mobilität in Städten (SrV)

2. Adresskoordinaten-feine Verteilung der Haushalte



Flächennutzungstypen
(Wohnbaufläche, Industrie- und Gewerbefläche, usw.)

Quelle:
Digitales Landschaftsmodell (DLM)
Bundesamts für Kartographie und Geodäsie (BKG)



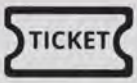
Gebäudedaten
(Wohnblöcke und Anzahl Bewohnern)

Quelle:
Amtliches Liegenschafts-katasterinformationssystem (ALKIS)

Adressdaten
(Koordinaten)

Quelle:
Adressdatensatz
Bundesamts für Kartographie und Geodäsie (BKG)

3. Ergänzung von Mobilitätsoptionen

Attribute	ca 20 Minuten	ca. 40 Minuten		ca. 25 Minuten	ca. 10 Minuten	ca. 85 Minuten
						
Alter	✓			✓	✓	✓
Geschlecht	✓			✓		
Status		✓1	✓2, 3			
HH-Größe	✓	✓		✓		
HH-Einkommen	✓	✓	✓	✓		✓
Anzahl Führerschein im HH	→	✓	✓	✓		
Anzahl Erwachsene im HH		✓	✓			
Anzahl Männer im HH (18+)		✓				
Alter älteste Person im HH		✓				
Anzahl Autos im HH			→	✓		✓
TAPAS HH-Typ						✓
Quelle	**MiD 2008**	**MiD/SrV 2008**		**MiD 2008**		**EVS 2003**

Status: 1)Anzahl Vollzeit Beschäftigte 2)Anzahl Teilzeit Beschäftigte 3)Anzahl Erwerbstätige

Attribute der erstellten synthetischen Bevölkerung

Grundbevölkerung				Erweiterte Bevölkerung	
Randsummen-kontrolliert		Nicht kontrolliert		Regressionsmodelle	
Alter	21	Kinder im Haushalt	0/1	Führerschein	0/1
Geschlecht	2	HH-Typ	18	ÖPNV-Abo	0/1
HH-Größe	5	HH-EK diskret		Anzahl Autos	0/1/2
HH-EK, gruppiert	6	*Bildungsstand*		Art der Autos	
Status	8	*Arbeitsplatzdetails*		Fahrrad	0/1
		Entfernung Arbeitsstätte		Verkehrszelle	
		Renten und Transferl.		Adresskoordinate	
		Nationalität		MB var	
		Migr.hintergrund		MB fix	
		Wohnungswechsel		MB ÖPNV	
		Zweitwohnsitz			
		...			

Herausforderung heterogene Datenbasis

- Synthese benötigter Attribute aus unterschiedlichen Datenquellen
  - Basisbevölkerung, Fortschreibungsinformationen, Mobilitätsoptionen, finanz. Situation
- Datenschutzanforderungen
- Geringe Fallzahlen für spezifische Untersuchungsräume
- Geringe räumliche Auflösung, insb. für Prognose
- Heterogene Attributsausprägungen, Bezugsjahre, räumliche Bezüge und Auflösungen, Angaben / Werte
- Langandauernde Datenakquisen, teilweise häufige Aktualisierungen

→ **Notwendigkeit umfangreicher Harmonisierungsarbeiten !**

Qualitätsprüfung während und nach der Erstellung

- Basisbevölkerung:
  - Randsummenprüfung während der Erstellung mit dem Synthesizer: Standardized Root Mean Square Error (SRMSE)[1]
  - Visuelle Prüfung relativer und absoluter Abweichungen (Karten, Verteilungen,…)
- Mobilitätsoptionen
  - Visuelle Prüfung relativer und absoluter Abweichungen (Karten, Verteilungen,…)

Personenverteilung
Alter (21) x Geschlecht (2) x Status (8)

$$SRMSE = \frac{\sqrt{\frac{\sum_{ijk}(Observed_{ijk} - Expected_{ijk})^2}{ijk}}}{\frac{\sum_{ijk} Expected_{ijk}}{ijk}}$$

SRMSE

$$\begin{cases} \min 0 \text{ (stimmt exakt überein)} \\ \max 1 \text{ (stimmt nicht überein)} \end{cases}$$

1 Knudsen & Fotheringham 1986 nach Prichard & Miller 201

Aktualität und Beschränkungen

- Separate Erstellung für jede betrachtete Zeitscheibe
- Aktualisierung der synthetischen Bevölkerung je nach Projektanforderungen:
  - Datenaktualität
  - Konsistenz mit anderen Eingangsdaten
  - spezifische räumliche Bezüge
  - spezifische Anforderungen an die Flotteninformationen

- Beschränkungen
  - Räumliche Auflösung der Daten relativ grob
  - Teilweise umfangreiche Harmonisierung notwendig
  - Fortschreibung und Prognose: Annahme Verhaltens- und Raumkonsistenz
  - SALSA: zahlreiche Attribute nicht vorhanden (Umzugshistorie, Zuzugsinformationen, Migrationshintergrund, Präferenzstruktur, …)

Quo vadis?

- Stärkere Nutzung der Methodik zur Fortschreibung von Strukturdaten in den Modellen
  - Unternehmensstandortwahl (SALSA)
  - Zielwahl (TAPAS)
- Überarbeitung und Aktualisierung der Modelle für die Mobilitätsoptionen gemäß neuer Verhaltensdaten (MiD und SrV 2017, EVS)
- Verbesserung der Fortschreibung, eventuell mit Evolutionsmodellen

- Nutzung auch separat von den beiden originären Ziel-Modellen
  - Analysen zur Ladeinfrastruktur und Pkw-Besitz
  - Mikrosimulationsmodell städtischer Güterverkehr
  - …

Heureka-Beitrag: elib.dlr.de/111905

Rita Cyganski
Institut für Verkehrsforschung
Rita.Cyganski@dlr.de
+49 (30) 670 55 – 147
www.dlr.de/vf

Kanton Zürich
Direktion der Justiz und des Innern
Statistisches Amt

Qualitätssicherung von synthetischen Populationen - Ein Erfahrungsbericht

Dr. Peter Moser, stv. Amtschef, Leiter Analyseabteilung

Ausgangslage

Das **Statistische Amt des Kantons Zürich (STAT)** begleitete, zusammen mit einem weiteren kantonalen Stakeholder, dem **Amt für Verkehr (AfV)**, die Erarbeitung der **synthetischen Population von EBP** während rund eines Jahres.

Unsere Mitarbeit bezog sich konkret auf:

- **die Teilschritte des Produktionsprozesses:** die Qualität, bzw. Plausibilität des Regelwerks der Aufbereitung von Inputdatensätzen aus Stichprobenerhebungen wie dem Mikrozensus Verkehr, oder der HABE und deren Verknüpfung mit den kleinräumig verfügbaren Registerdaten (STATPOP, GWS).
- **Die Plausibilität des Resultats** verglichen mit der "Wirklichkeit", bzw. den aus anderen Datenquellen bekannten stilisierten Fakten der räumlichen Verteilung von Merkmalen.

Bsp. 1: Qualitätssicherung Prozess

Problemstellung: Schätzung einer hypothetischen **individuellen Jahresmobilität nach Hauptverkehrsmitteln** für die Zielpersonen des MZVM auf der Basis tatsächlich erfragter **individueller Tagesdistanzen**.

Komplexer Imputationsprozess (Grundsatz: nur tatsächlich vorkommende Tagesmuster sollen verwendet werden), wird durch zahlreiche Annahmen gesteuert, die im einzelnen zu hinterfragen sind:

- **Soziodemographische Gruppierung** (Wohnort, Haushaltstyp, Alter): Ist die Klassenbildung plausibel, sind die Gruppen hinsichtlich ihres Verkehrsverhaltens in sich homogener als zwischeneinander?
- **Annahmen hinsichtlich durchschnittlicher Jahresfahrleistungen** von Auto- und ÖV-Abonnementsbesitzern im Werktagsverkehr: Sind sie plausibel?

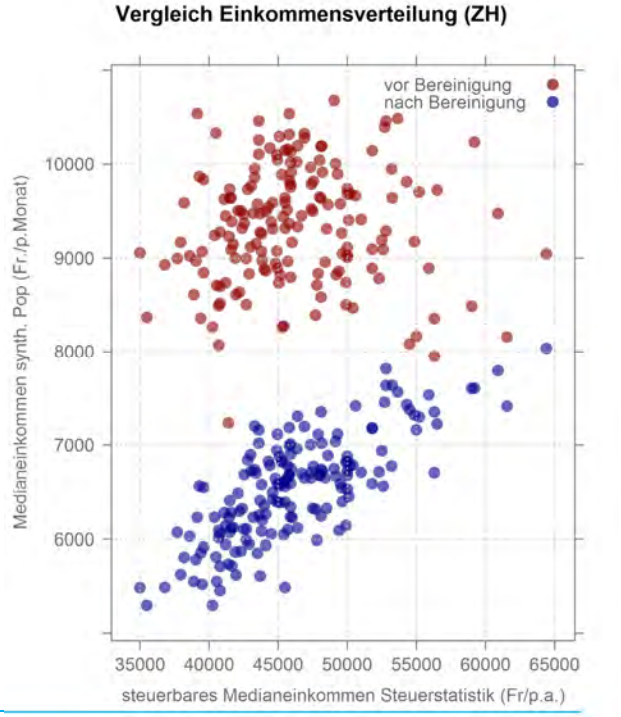
➔ Vergleich der mittleren Jahresmobilität nach Verkehrsträgern zwischen MZVM-Originaldaten und Imputat auf gesamtschweizerischer Aggregatsebene ergab eine erhebliche Überschätzung (Faktor 2) der ÖV und LV-Jahreskilometer. **Die Annahmen mussten entsprechend angepasst werden**.

Bsp. 2: Qualitätssicherung Resultat

Problemstellung: Gibt die synthetische Population die "stilisierten Fakten" der räumlichen Verteilung der modellierten Merkmale grundsätzlich korrekt wieder?

- **Entspricht die räumliche Verteilung der mittleren Einkommenshöhe (**Quelle: HABE) **in der synthetischen Population dem bekannten Muster der Steuerstatistik?**

➔ In einer ersten Version (vor Bereinigung) war dies nicht der Fall. Eine **Nachbearbeitung der Zuordnungsregeln brachte erhebliche Verbesserungen** (nach Bereinigung)



Möglichkeiten und Grenzen der Qualitätssicherung

Stufe 1: Formal-Technischer Aspekt - wurde korrekt gerechnet?

- **Systematischer Vergleich von Aggregaten** (Häufigkeiten, Kennwerte, wie Mittelwert und Streuung) **der Originaldatensätze und der synthetischen Population** (oder von Zwischenprodukten, wie der Jahresmobilität im MZVM) auf der räumlichen Ebene, die der Modellierung zugrunde liegt.

Stufe 2: Inhaltliche Aspekte i.e.S. - ist ein Mehrwert vorhanden?

- **Eine synthetische Population ist dann "gut"**, wenn sie – im Schnitt und unter Bewahrung der meist erheblichen Variabilität – für die **Individuen/Haushalte plausible Merkmalsbündel** erzeugt: Wohnort/-situation, finanzielle Verhältnisse, Mobilitätswerkzeuge und Verkehrsverhalten sollten in Ausmass und Vorzeichen "wirklichkeitsgetreu" miteinander zusammenhängen. **Dann stimmen auch die Korrelationen mit relevanten Indikatoren auf Aggregatsebene.**
- **Für diesen "Realitätsbezug" lässt sich wohl kein simples Testprotokoll mit definierten Messgrössen und klaren Kriterien formulieren: Kreativität, inhaltliches Fachwissen ist gefragt - und nicht zuletzt (kostbare!) Zeit**.

Fazit - Anregungen

- **Stufe 1: Überprüfung der formal-mathematische Korrektheit des Imputationsprozesses** (RV Stichprobenerhebung = RV fiktive Vollerhebung) **sollte festes Element der Produktion sein** (allenfalls mit Toleranzgrenzen!)
- **Stufe 2: Plausibilitätschecks i. e. S. könnten strukturiert werden durch**:
  - **Definition räumlicher Aggregationsniveaus** für Korrelationsanalysen (Region, Gemeinde, kleinräumigere Aggregate).
  - **Beizug thematisch verwandter Datensätze, welche nicht in die Imputation einfliessen**, aber auch **Informationen zu interessierenden Grössen** enthalten: z. B. Strukturerhebung (einkommensrelevante Bildung, berufliche Stellung, Pendlerverhalten).
  - **Vorgängige Vereinbarung der "stilisierten Fakten"**, die sich in der synthetischen Population spiegeln sollen. Hilfreich könnte hierbei auch der Einbezug externer Experten mit fachübergreifendem Wissen sein.
- **Nicht zuletzt: Der testende Nutzer muss das Produkt auch in einer adäquaten, für ihn handhabbaren Form erhalten** (z.B. vermittelt durch eine einfache Applikation?)

Bedürfnisse der Bundesverwaltung & Einsatz in den Themen Raum und Verkehr

Bern, 08.12.2017

Andreas Justen, Nicole Mathys
ARE Sektion Grundlagen

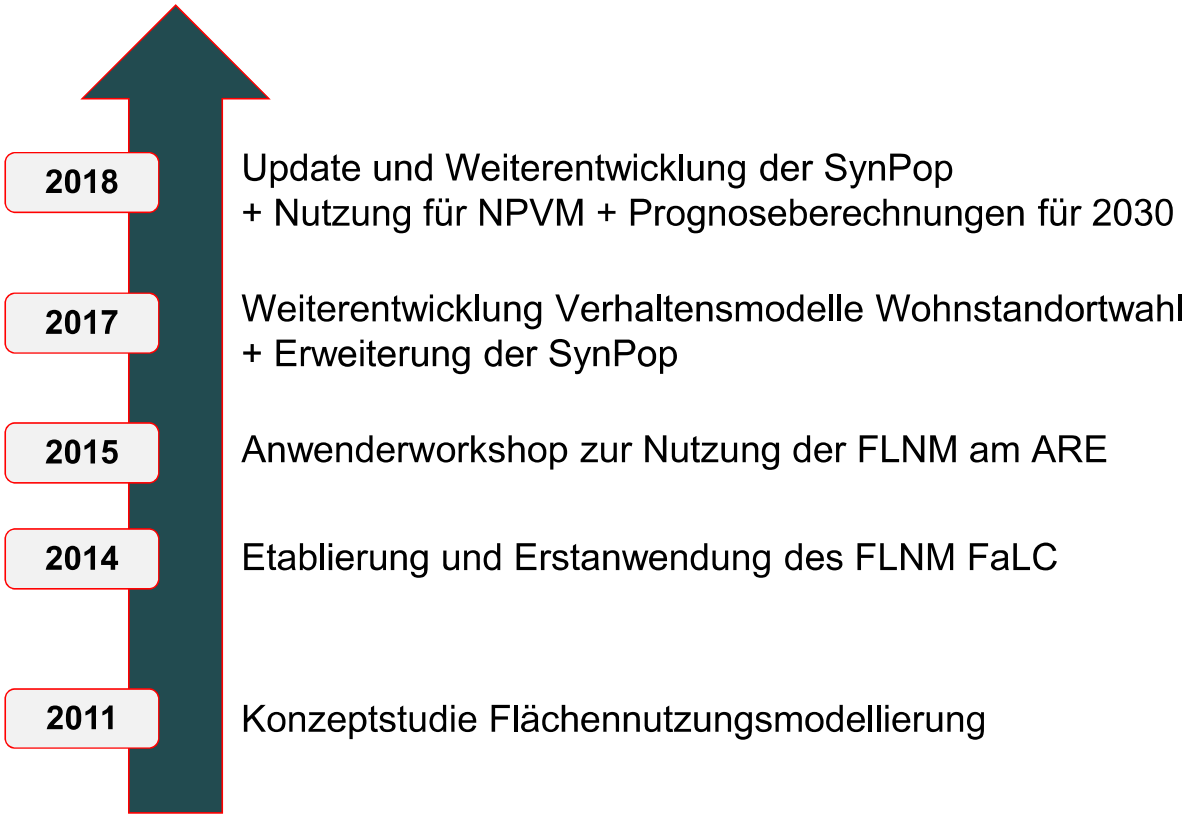
Agenda

- Projekthistorie am ARE
- Einsatzfelder SynPop bzw. Flächennutzungsmodell (FLNM)
- Anforderungen an SynPop seitens Bundesverwaltung und ARE
- Leistungen des ARE

Ansätze für die Diskussion

- Einsatz NPVM + FLNM
  (Integrierte Modellierung von Verkehrs- und Raumentwicklung)
- Ausblick Einsatzfelder SynPop

Projekthistorie SynPop / Flächennutzungsmodellierung



Einsatz in den Bereichen Raum & Verkehr
→ heute und in Zukunft

- **Nationales Personenverkehrsmodell (NPVM) / Neu bis 2019**
  - Ableitung der 81 verhaltenshomogenen Personengruppen
  - Personen nach: Alter, Mobilitätswerkzeugbesitz, Raumtyp

- **Flächennutzungsmodell / Entwicklung und Anwendung**
  - Modellierung der Wohn- und Unternehmensstandortwahl
    → Bevölkerungs- und Arbeitsplatzverteilung
  - Prognose der SynPop 2030/2040 & Verwendung im NPVM

Eigenschaften der «ARE-SynPop»
→ Attribute im NPVM

- Person: **Alter**, Geschlecht
- Haushaltstypen (+ Rolle)
- Schüler, Studierende
- Eigentümer (ja/nein)
- **Erwerbstätig** (ja/nein)
- Arbeitspensum
- CH, Ausländer
- Bildungsniveau
- Nationalität
- Arbeitsplatz (nach Branche)
- **PW-Besitz & Verfügbarkeit**
- Einkommen (Person, Haushalt)
- **Abo-Besitz ÖV**

Anforderungen seitens Bundesverwaltung / ARE

- Verwendung aktueller Grundlagen von STATPOP / STATENT;
- Erweiterung mit (bestmöglich) verfügbaren, weiteren Daten: Strukturerhebung, Lohnrechner, Haushaltsbudgeterhebung, MZMV;
- Einsatz von Ergebnissen statistischer Verhaltensmodelle, v.a. im Bereich der Wohnstandortwahl und Wahl von Mobilitätswerkzeugen *(je nach Thematik andere Verhaltensmodelle nötig)*;
- Berücksichtigung von Prognosevorgaben bei Personen und Haushalten (= BFS Szenarien zur Bevölkerungs- und Haushaltsentwicklung);
- Nachweis der Validierung entlang zentraler Variablen.

Erwartungen des ARE / FLNM mit explizitem Raumbezug

Validierung und Kalibration auf jeweils tiefster Raumebene

- Bevölkerung (Altersgruppen) und Arbeitsplätze (Branchen)
  → **Adresse, Hektare, Verkehrszone**
- Besitz Mobilitätswerkzeuge (PW, GA, HTA, VA)
  → **Verkehrszone**
- PW-Verfügbarkeit
  → **Kanton gekreuzt mit Gemeindetyp, *(MS-Regionen)***
- Verfügbares Einkommen (Netto)
  → **CH, Kanton, *(Gemeinden)***
- Gekreuzte Eigenschaften (z.B. Haushaltstyp + Anzahl Mobilitätswerkzeuge + Einkommen)
  → **CH, Kanton gekreuzt mit Gemeindetyp**

Exkurs: Verkehrszonen NPVM

Schweiz: Neu von 2'944 auf **7'978** Zonen.
Im Mittel **1'660** Einwohner + Vollzeitäquivalente pro Zone.
Aggregierbar zu allen CH-Gemeindeständen ab dem Jahr 2000.

Anzahl Verkehrszonen in 5 grössten CH-Städten

	Anzahl
Basel	141
Bern	105
Genève	126
Lausanne	88
Zürich	308
Summe	**768**

Leistungen des ARE

Bereitstellung bestmöglicher Grundlagen:

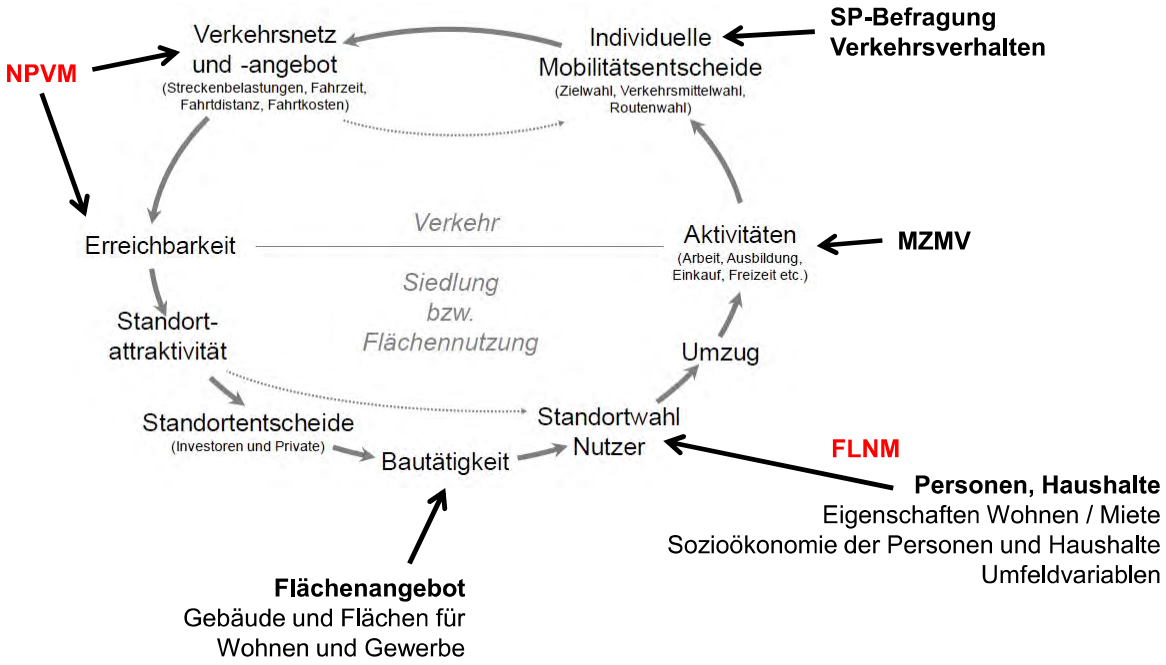
- BFS-Quellen: Anzahl Schüler, Studierende geokodiert (Weitergabe auf Stufe Verkehrszonen)
- Mobilitätswerkzeuge (Anzahl PW, GA, HTA, VA) geokodiert (Weitergabe auf Stufe Verkehrszonen)

Interaktion Modelle Raum & Verkehr / NPVM & FLNM

Warum ist eine Interaktion wünschenswert?

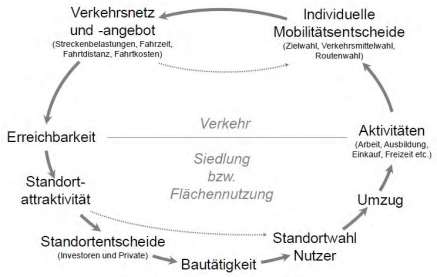
- Umzugsentscheide u.a. abhängig von der Erreichbarkeit der Wohn- und Arbeitsplatzstandorte;
- Verteilung von Bevölkerung und Arbeitsplätzen wirkt in Verkehrserzeugung sowie Zielwahl im Verkehrsmodell.

Interaktion Modelle Raum & Verkehr / NPVM & FLNM



Modellierung Raum & Verkehr

→ *Umsetzung Verkehrsperspektiven 2040*

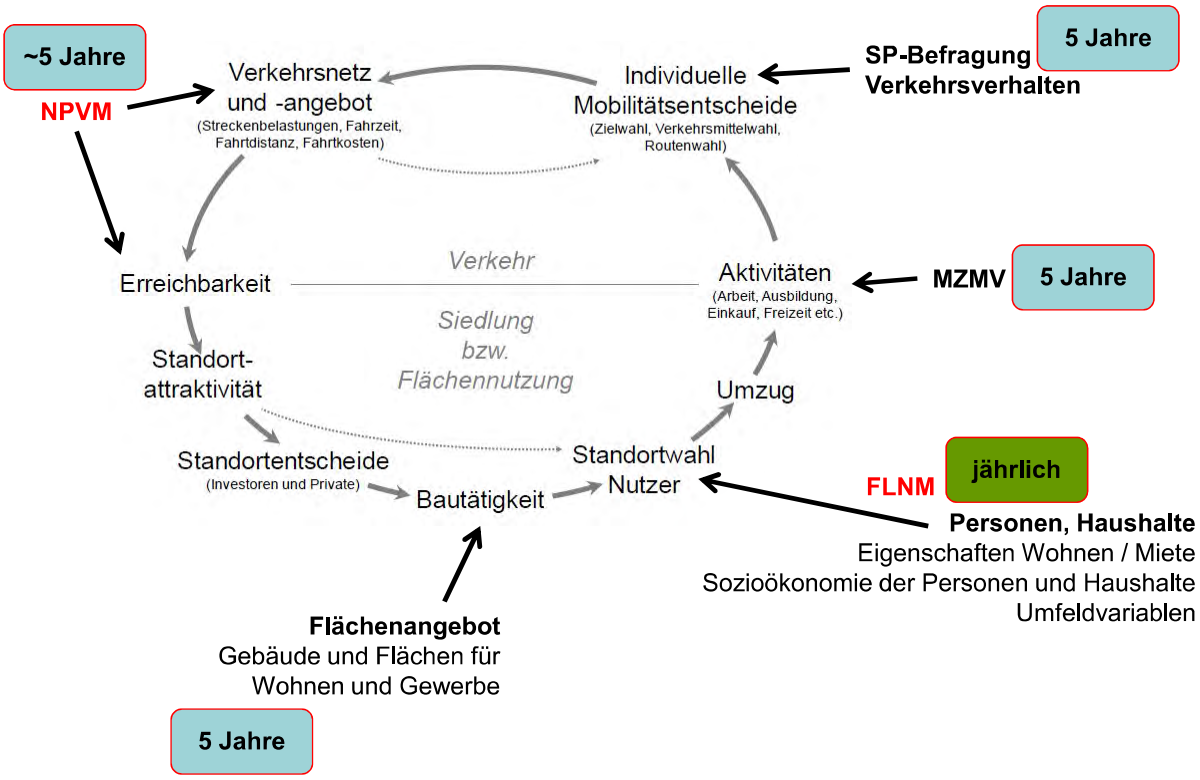


- Erreichbarkeiten 2010 / 2020 / 2030 / 2040:
  - Entwicklung Infrastruktur Strasse für die 4 Zeitpunkte abbildbar (ASTRA-Projekte)
  - Entwicklung Fahrplan ÖV 2010 sowie STEP 2025 gültig für 2020, 2030 und 2040
  - **Faktormatrix Strasse**: BEL_2010 / UNBEL_2010
    → Anwendung auf die UNBEL_2020/30/40
  - Annahme: *lineare*, räumliche Entwicklung von Bevölkerung und Arbeitsplätzen

- Bevölkerungs- und Arbeitsplatzentwicklung
  a) Arbeitsplätze (u.a. in Abhängigkeit von Erreichbarkeiten Strasse & ÖV, siehe oben)
  b) Bevölkerung in Abhängigkeit von a) und Erreichbarkeiten Strasse & ÖV, siehe oben
  c) → Modellierung der effektiven Verkehrsnachfrage in 2020, 2030 und 2040

- Interaktion Raum / Verkehr nur näherungsweise abgebildet

Interaktion Modelle Raum & Verkehr / NPVM & FLNM



Einsatz in den Bereichen Raum & Verkehr
→ heute und in Zukunft

- **Beispiele für die weitergehende Verwertung des «Produkts» Synthetische Population**
  - Ist-Nachfrage nach Mobilität (bezogen auf den Wohnstandort), z.B. über Kopplung mit Kenngrössen aus dem MZMV;
  - Analysen zum Zusammenhang zwischen Mobilitätswerkzeugbesitz und Erschliessungsqualitäten;
  - Analysen zur Ist-Nachfrage nach Wohnformen (+ Ableitung der Bedeutung für neu entwickelte Quartiere);
  - Analysen zur Ist-Nachfrage im Energiebereich (Mobilität, Wohnen);
  - …