Simulation based Population Synthesis

Work in progress…

Bilal Farooq
Michel Bierlaire
Gunnar Flötteröd

Seventh Workshop on Discrete Choice Models

August 25–27, 2011

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Contents

- Current state-of-the-art
- Problem statement
- Methodology
- Implementation
- Initial results
- Conclusion and future direction

Context

- Microsimulation: Forecasting behaviour using individual level models
  - Lack of individual level data for population
  - Synthesis of individual agents and their characteristics
- Initial work
  - Focused on synthesis of a small sub-set of characteristics
    - Usage: Activity-base travel demand models, etc.
- Frequently used approaches
  - Fitting based
    - Combinatorial optimization
    - Iterative proportional fitting

Definition

- Synthetic Agents
  - Households, persons, families and their association
- Space/location
  - Zone: parcel/sector/commune, dissemination area/tract/sub-division, traffic analysis zone/planning district
- Time
  - Base year, *simulation year*
- Data
  - Aggregate: zonal level totals and cross-tables
  - Disaggregate: sample of individuals, usually without location [Public Use Microdata Sample (PUMS)]
    - Public Use Microdata Area (PUMA)

Literature Review

- Combinatorial optimization (Williamson, 1998)
  - Zone-by-zone
  - Integer weights for each row in the PUMS
    - 0 or 1 for an observation in the PUMS, to be included in the zone (note, no duplication of obs. within the zone)
  - Use of hill-climbing, simulated annealing, and genetic algorithm to estimate the best set of obs. weights for each zone

Literature Review

- CO approach strengths
  - Efficient in terms of memory (compared to IPF)
  - Continuous/discrete characteristics
  - Adding layers of different data is trivial
  - Low variance (Huang Williamson, 2001)
- Weaknesses
  - Correlation structure (association patterns) within the zonal population is not explicitly conserved
    - PUMA level constraints needed to be applied
  - Extremely large search space

Literature Review

- Iterative Proportional Fitting (IPF) [Beckman et al. 1996]
  - Basic idea
    - Steps: fitting, generation
    - Contingency table
      - Categorization of variables of interest
      - Totals for each cell of the resulting multi-way table
    - Multi-constraint gravity model sort of formulation
      - Iterate while the error is large
      - Use marginal as dimensional totals
      - In each step
        - Adjust the cell probabilities to fit dimension totals

Literature Review

- Iterative proportional fitting
  - Basic idea
    - Odd ratio concept (ratio between any two cells before and after fitting remains the same)
    - Zone-by-zone vs Multi-zone IPF
    - Integerization of contingency table
      - Realization from the contingency distribution
      - Integer and fractional component can also be separated

Literature Review

- Iterative proportional fitting
  - Issues
    - Dimensionality
      - Hyper dimension tables
    - Zero cell problem
    - Integerization of cell values
    - Association structure for household and individuals
    - Higher variance
    - High degree of ad-hocness in the procedures

Literature Review

- Historic improvements in IPF
  - Beckman et al. (1996), Frick and Axhausen (2004), Arentze et al. (2007), Guo and Bhat (2007), Pritchard and Miller (2009), Ye (2009) IPU, Auld et al. (2010), and Barthelemy and Toint (2011)

Comments

- Key issues with the existing approaches
  - Cloning of data rather than creation of a heterogeneous representative population
  - Over reliance on the accuracy of the microdata, without serious consideration to the sampling process and assumptions
  - Optimization resulting in one realization of synthetic population
  - Scalability issues

Problem statement

- We are interested in building a joint distribution of the population from which one or more realizations of synthetic population can be created, such that
  - Representative of the real population
  - Synthetic population having a “continuous heterogeneity” rather than “discretized cloning”
  - Population synthesis process as a part of the microsimulation
  - Methodology does not need to know the data collection and aggregation process

Methodology

- Synthetic Agents: Persons (X), Households (H)
  - X and H defined by their characteristics
- Associations (C)
- Use simulation to construct $\pi(X, H, C)$
- Available Data
  - Data on persons characteristics Y_X
    - $\pi(X|Y_X)$
  - Data on households characteristics Y_H
    - $\pi(H|Y_H)$
  - Data on association characteristics Y_C
    - $\pi(C^i|Y_C,X^i,H^i)$

Methodology: Persons

- Persons synthesis (X)
  - Method: Gibbs sampling
    - Conditionals for person characteristics known to certain extent
    - Run a Markov Chain Monte Carlo simulation to generate the persons
    - Results in an infinite pool of feasible persons
    - Realization from this universe will result in the synthetic population of persons

Methodology: Persons

- X = {Age, Sex, Marital_Status, Dwell_Type}
- Data needed: Zone (sector, commune) level conditionals
  - E.g. ***P***(Age | Marital_status, Sex, Dwell_Type), ***P***(Marital_status | Age, Sex, Dwell_Type), ***P***(Sex | Age, Marital_status, Dwell_Type), ***P***(Dwell_Type | Age, Marital_status, Sex)
- Each iteration for X_t
  - Pick a characteristic and realize its value from its conditional based on the other characteristics of X_{t-1}

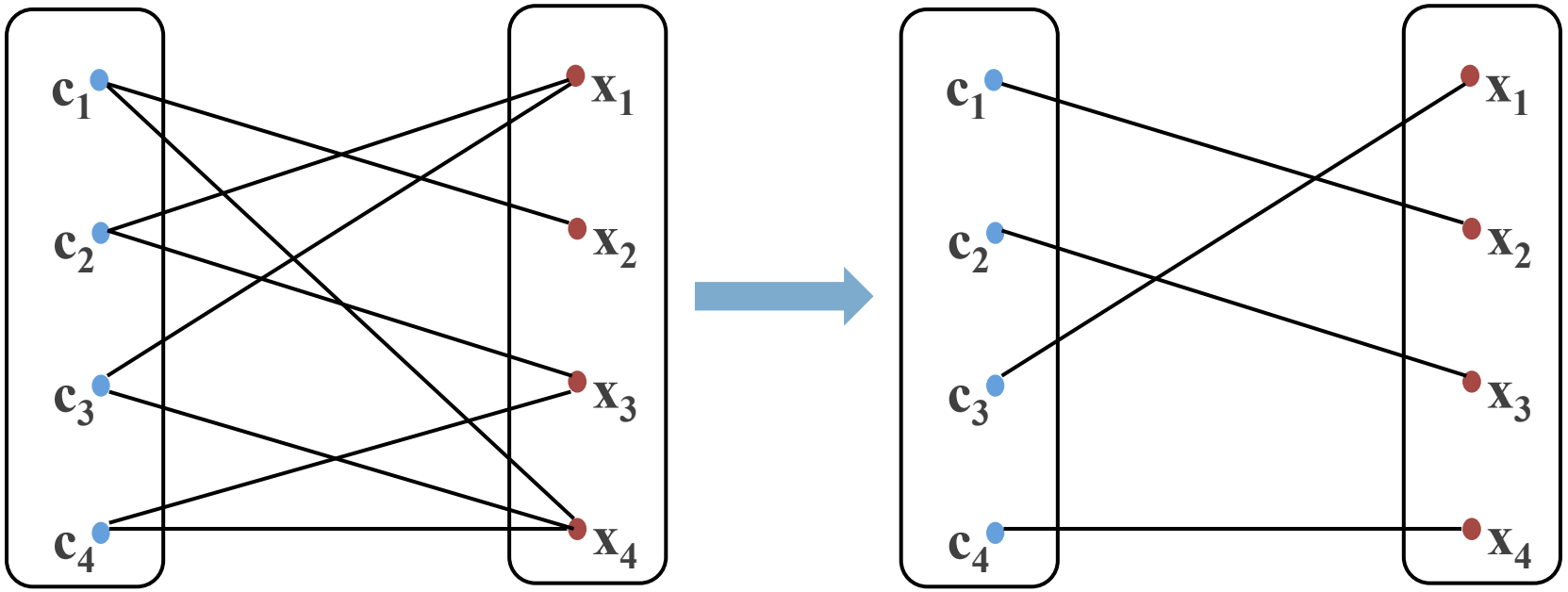
Methodology: Household

- Household synthesis (H)
  - Method: Gibbs sampling
  - Data needed: Zone (sector, commune) level conditionals
    - E.g. ***P***(Hhld_Type | Veh_Count, Dwell_Tenure, Dwell_Type), ***P***(Veh_Count | Hhld_Type, Dwell_Tenure, Dwell_Type), ***P***(Dwell_Tenure | Hhld_Type, Veh_Count, Dwell_Type), ***P***(Dwell_Type | Hhld_Type, Veh_Count, Dwell_Tenure)
  - Synthetic households as encapsulation of positions
    - From the realization of households
      - A list of available positions to be filled by persons

Methodology: Associations

- Associations (C)
  - Matching Persons to Positions
    - Head of household, housewife, children, adults
  - For each realization a distribution of association is computed that is based on the available microdata
    - Minimizing the difference in count with the microdata

Methodology: Association



Methodology

- Method: Metropolis-Hasting sampling
  - State: A valid assignment
    - Examples: a two years old is not the head of household
  - Proposal matrix/function
    - Defined in terms of switching the association of two persons with each other (bi-directional)
  - Initialized to certain random state of association
  - Transition/proposal distribution
    - Acceptance rate (awarding good states and penalizing bad)
  - A realization from the distribution

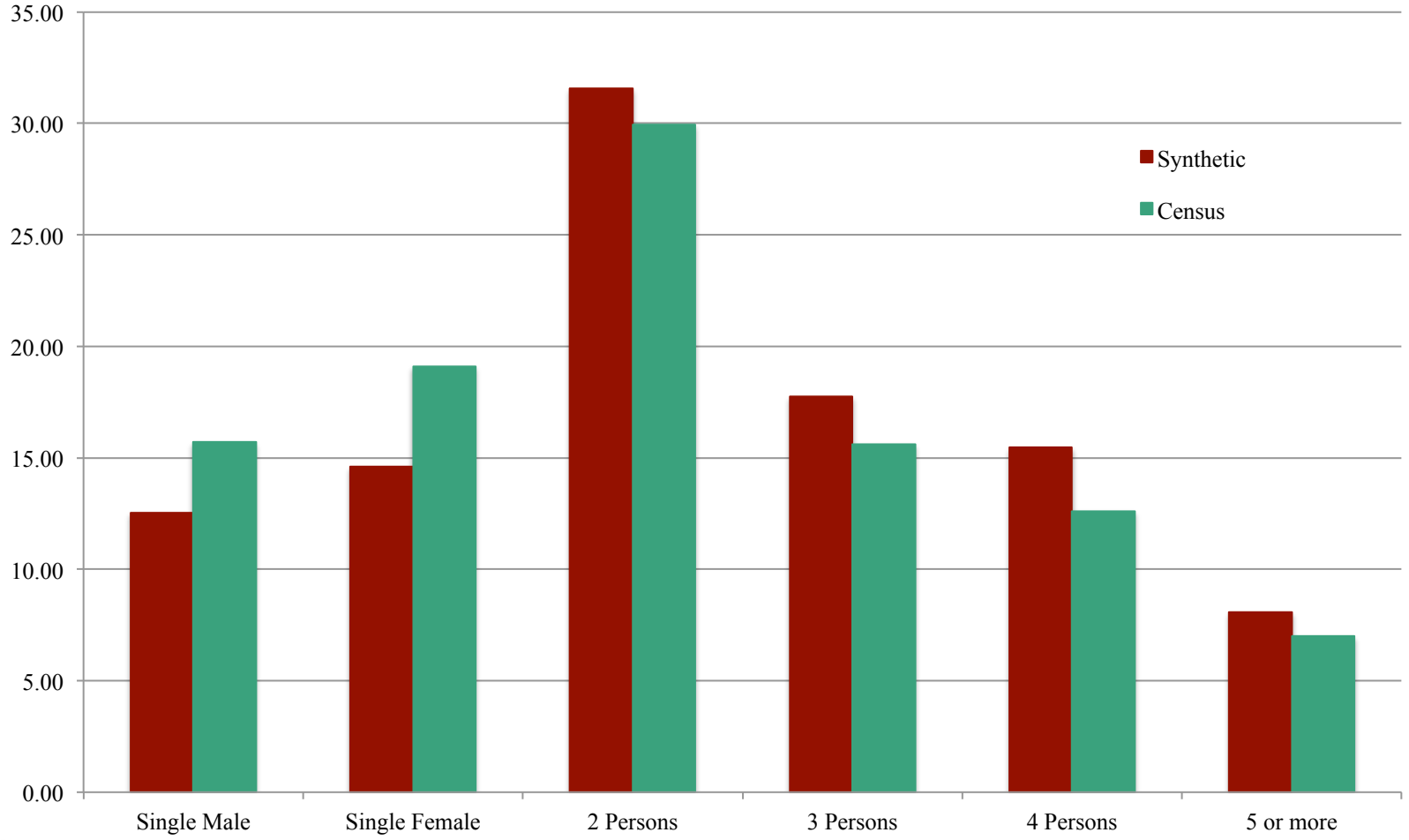
Implementation

- Prototype implemented in C^{++} for agents synthesis
- Synthetic population for Brussels region
  - Data sources
    - Census 2001 for zonal conditionals of households and persons
    - MOBEL 1999 travel survey of households and individuals

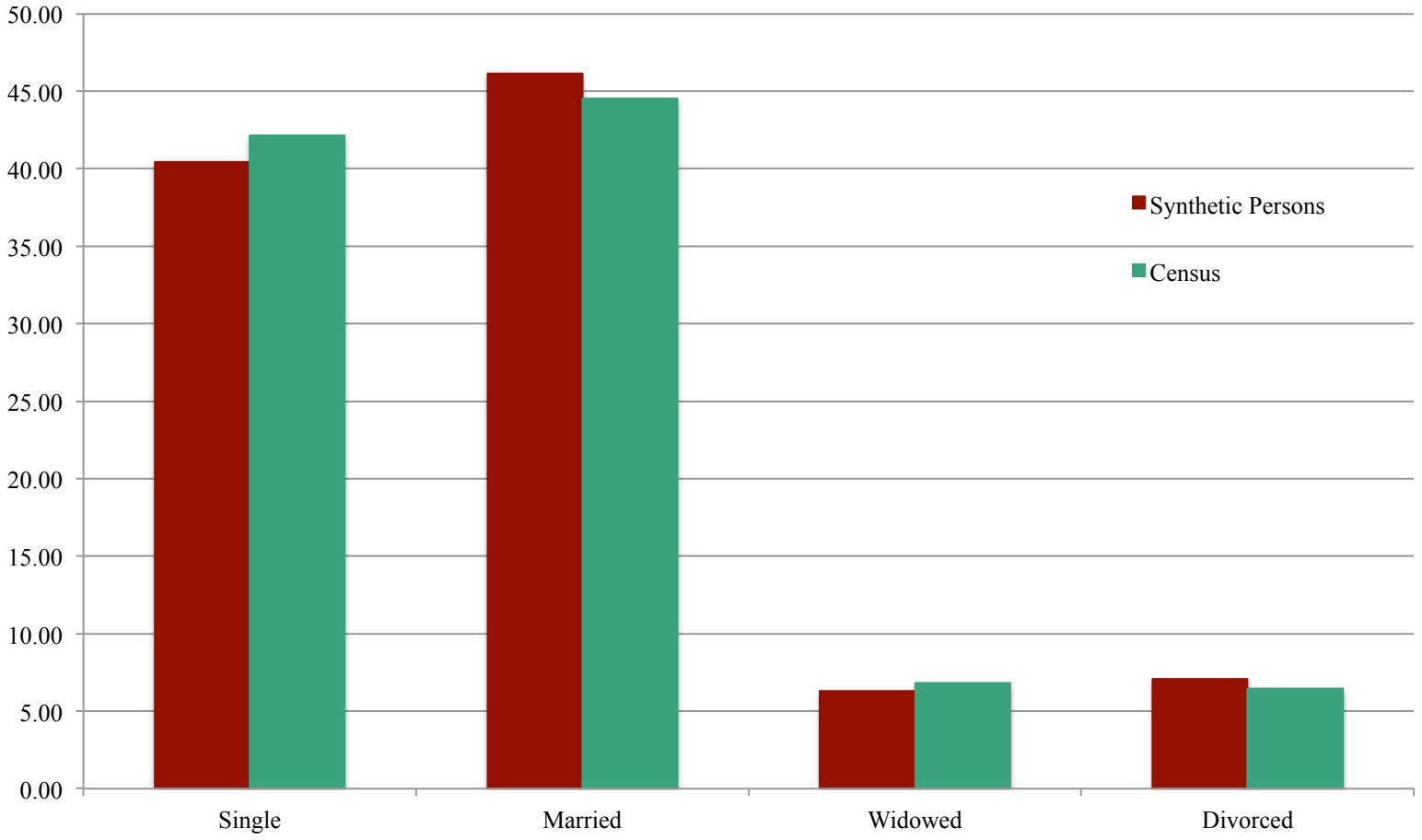
Progress to date (Results)

- Synthetic persons population
  - Attributes synthesized: marital status, sex, location (sector), nationality
    - Computational time < 1min
  - Attributes to be added
    - Ages, Income, education, dwelling
- Synthetic household population
  - Attributes synthesized: household size, vehicle fleet size, location (sector), tenure, dwelling type
    - Computational time < 1min

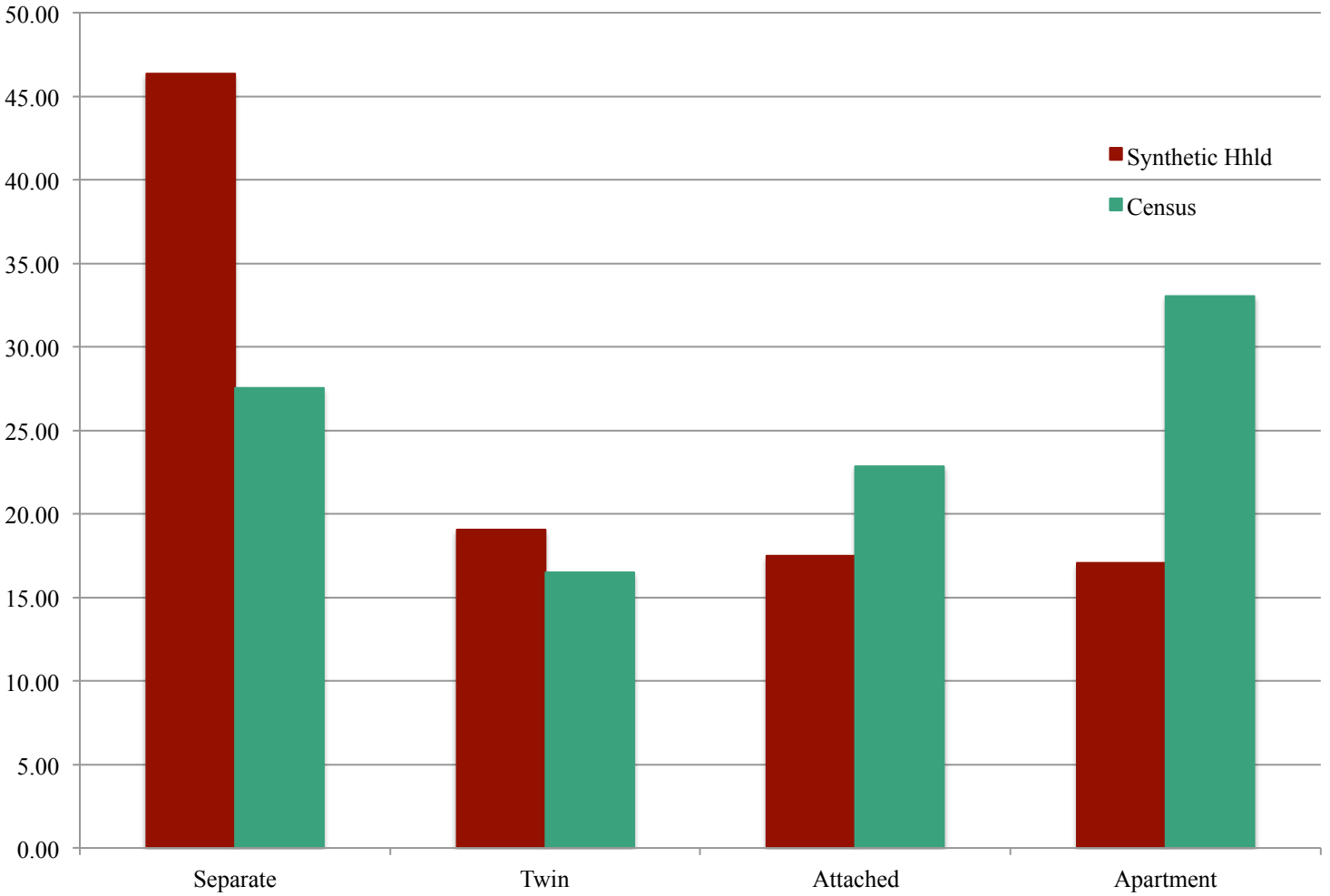
Household Size (Brussels, 2001)



Persons' Marital Status (Brussels, 2001)



Households' Dwelling (Brussels, 2001)



TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Synthetic Persons (Brussels, 2001)

Sex	Synthetic	Census
Male	48.69	48.61
Female	51.31	51.39

Martital Status	Synthetic	Census
Single	40.43	42.17
Married	46.16	44.55
Widowed	6.32	6.83
Divorced	7.09	6.44

Nationality	Synthetic	Census
Belge	91.49	88.44
Etrange	8.51	11.56

Aggregate level Percentages

Synthetic Households (Brussels, 2001)

Household Size	Synthetic	Census
Single Male	12.55	15.71
Single Female	14.59	19.11
2 Persons	31.56	29.96
3 Persons	17.75	15.62
4 Persons	15.47	12.61
5 or more	8.08	7.00

Vehicle Fleet	Synthetic	Census
None	18.35	25.13
1	51.87	52.07
2	26.39	20.53
3 or more	3.39	2.28

Dwelling Tenure	Synthetic	Census
Rent	24.37	34.01
Own	75.63	65.99

Dwelling Type	Synthetic	Census
Separate	46.35	27.55
Twin	19.08	16.51
Attached	17.48	22.88
Apartment	17.09	33.06

Aggregate level Percentages

Conclusions and Direction

- Proof of concept that the we are able to achieve the agents level stationary distribution
- Works both for continuous and discrete or mixture of conditionals
- Computationally efficient and scalable
  - Clean and simple
- Completion of association mechanism
- More detailed disaggregate level spatial and statistical analysis
- Using Swiss census to compare this with other approaches

Merci