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

1. Motivation
2. New methodology
3. Comparative experiments
4. Back to original problem
5. Concluding remarks

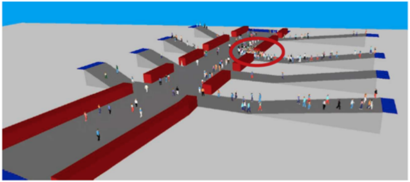
Modelling and Micosimulation

Urban area

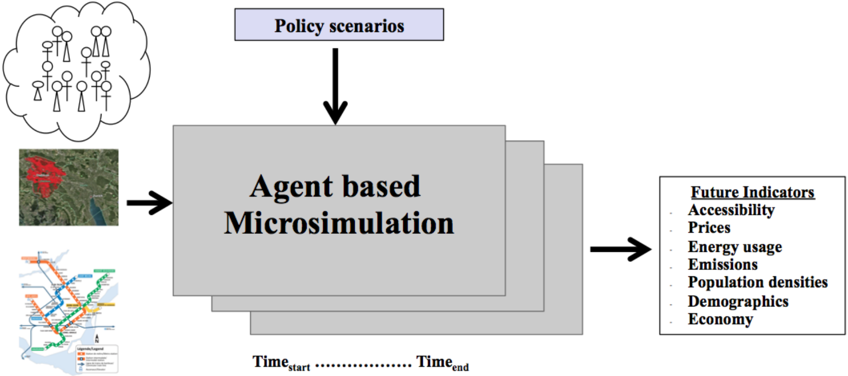
$$\Pi = \sum_{i=1}^{N} \frac{\gamma_i}{\alpha_i}\left\{(f^r(X_i^r) - f^c(X_i^c))^{\theta}\left(\left(\frac{q_i}{\gamma_i}+1\right)^{\alpha_i} - 1\right)\right\} + f^z(z)$$

Mobility Hub

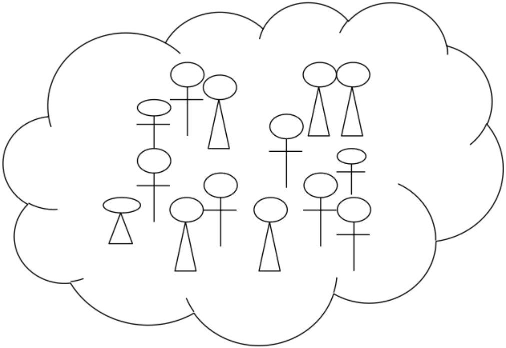
$$Q(\xi,\tau) = \begin{cases} n(\xi,\tau)\left\{1 - \exp\left[-\gamma_\xi A_\xi \left(\frac{1}{n(\xi,\tau)} - \frac{1}{N_\xi}\right)\right]\right\} & \text{if } 0 < n(\xi,\tau) < N_\xi \\ 0 & \text{otherwise} \end{cases}$$



Agent based Microsimulation



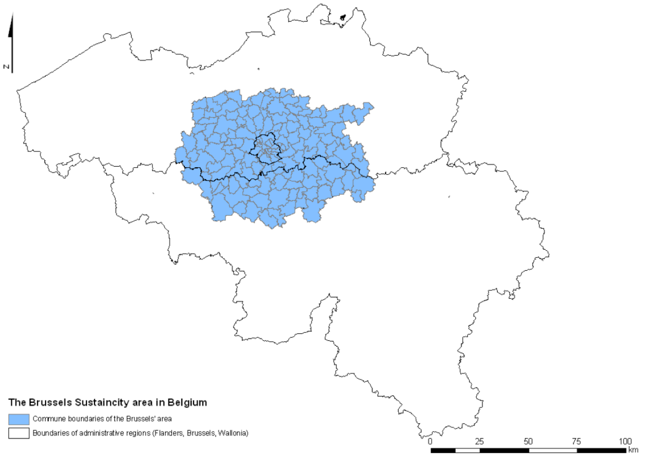
Population Synthesis



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 $\pi(X^1|X^{-1}) = \pi(X^1)$

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

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*

TRANSP-OR

EPFL ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Completing conditionals by domain knowledge

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

Completing conditionals by domain knowledge

- In case of domain knowledge
  $\pi(X^1|X^{(2\ldots k)}, X^{((k+1)\ldots n)} = a) = \pi^a(X^1|X^{(2\ldots k)})$,
  $\pi(X^1|X^{(2\ldots k)}, X^{((k+1)\ldots n)} = b) = \pi^b(X^1|X^{(2\ldots k)})$,
  ...
- 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 $Age = 1 \ldots 12$
    - $(Income|Sex, Age) = \beta(Income|Sex)$ for $Age = 13 \ldots 18$
    - $(Income|Sex, Age) = \gamma(Income|Sex)$ 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 $\pi(X_l^1|X_m^{-1}) = \dfrac{e^{(V_{X_l^1}|X_m^{-1})}}{\sum_{p=1}^{L}\left(e^{(V_{X_p^1}|X_m^{-1})}\right)}$

- 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

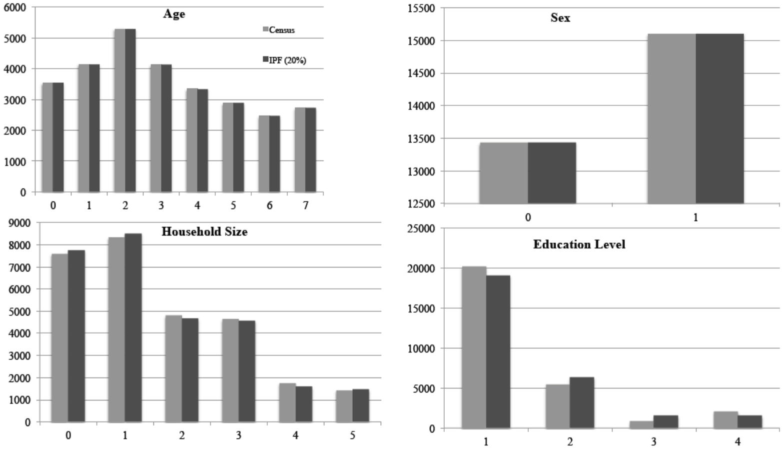
List of available conditionals

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

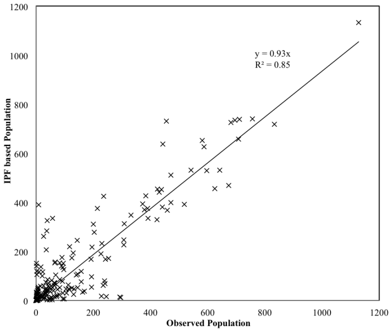
Data preparation

- 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

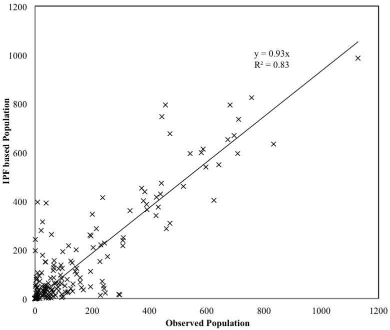
Results: IPF and Census marginals



Results: Fit of IPF with Census joint distribution



IPF with 20% sample

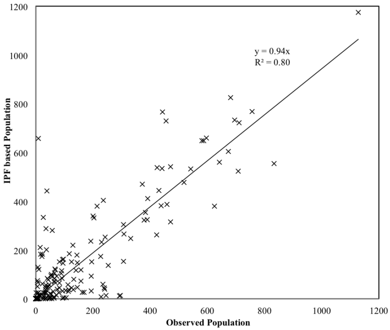


IPF with 10% sample

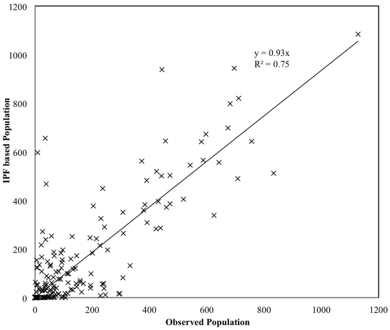
TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Results: Fit of IPF with Census joint distribution

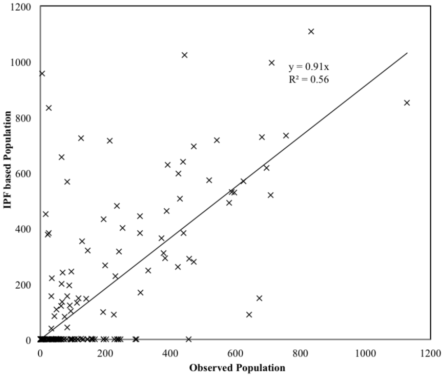


IPF with 5% sample



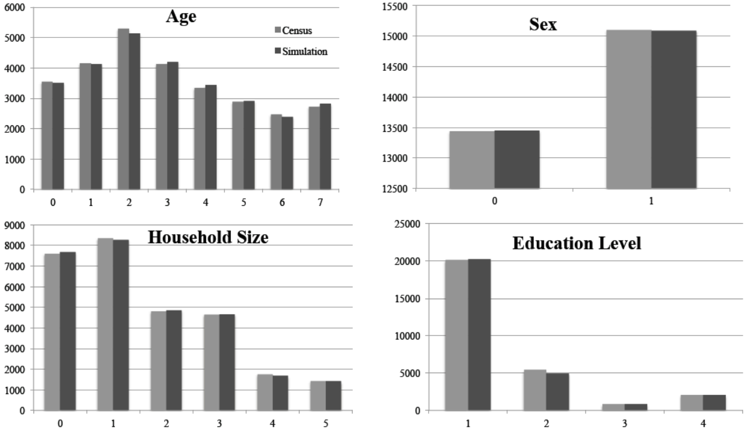
IPF with 3% sample

Results: Fit of IPF with Census joint distribution



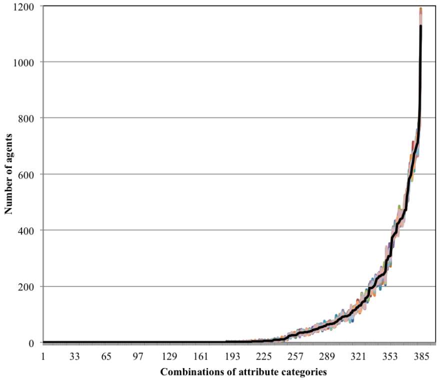
IPF with 1% sample

Results: Simulation and Census marginals



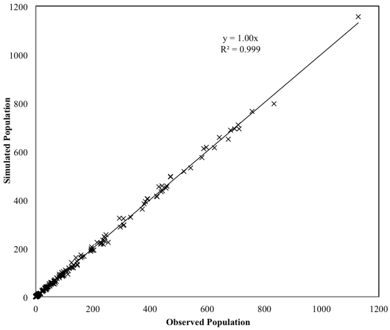
Using full-conditionals (*FullCond*)

Results: Simulation and Census joint dist.

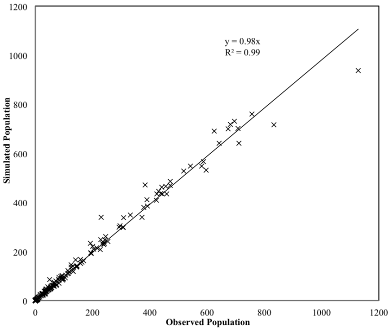


20 runs based on *FullCond* with real population superimposed

Results: Fit of Simulation with Census joint dist.

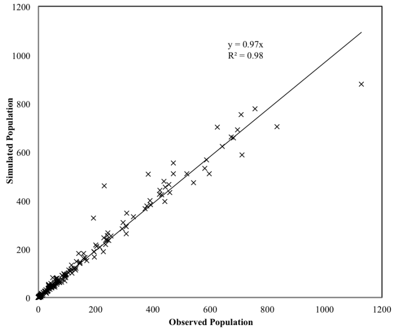


FullCond

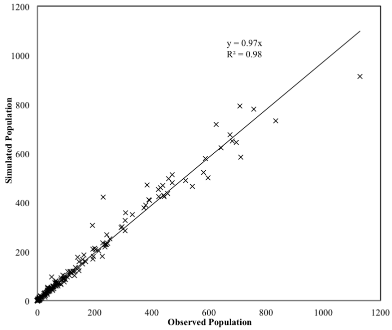


Partial_1 (Sex missing in 1 conditional)

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

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

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

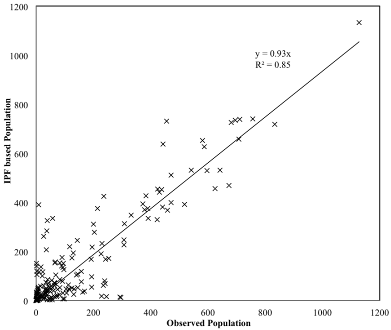
Comparison: Standard Root Mean Square Error

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

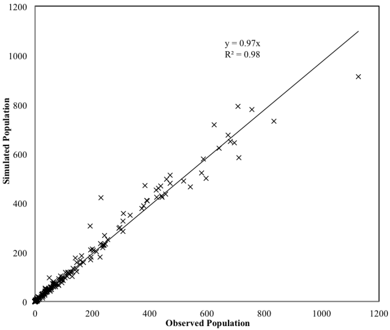
Input	IPF	Simulation
20% *Sample*	0.853	-
10% *Sample*	0.928	-
5% *Sample*	1.020	-
3% *Sample*	1.160	-
1% *Sample*	1.730	-
FullCond	-	0.130
*Partial*_1	-	0.240
*Partial*_2	-	0.340
*Partial*_3	-	0.350

- For Marginals only, both methods give the same fit

Best case IPF and worst case Simulation

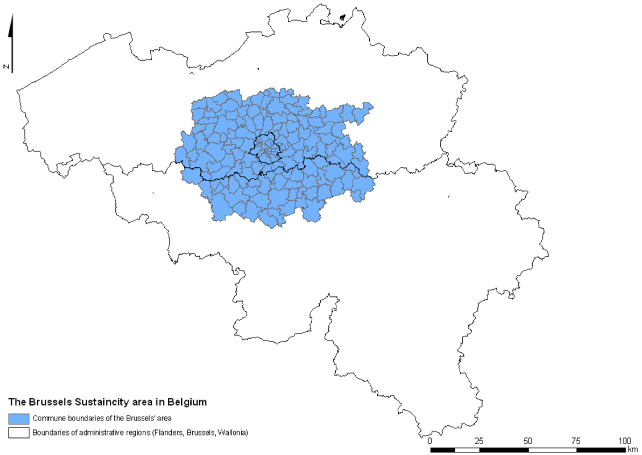


IPF with 20% sample



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

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)

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

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

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

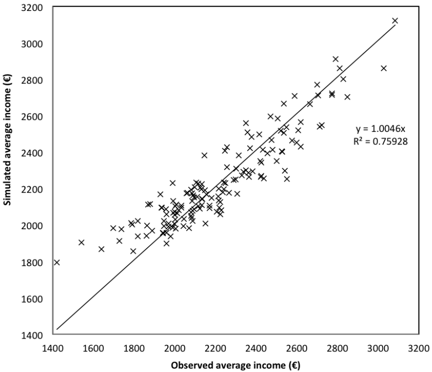
$$\begin{aligned} V^4_{(hh,z)} = ASC^4 + \beta^4_{educ_{hh}} \times educ_{hh} + \beta^4_{zonal\_inc_z} \times zonal\_inc_z + \beta^4_{cars_{hh}} \times cars_{hh} \\ + \beta^4_{house_{hh}} \times house_{hh} + \beta^4_{workers_{hh}} \times workers_{hh} \end{aligned}$$

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

Income level model

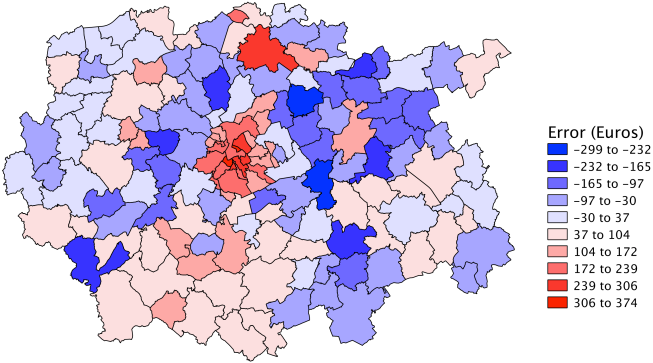
Parameter	Variable	Value	Std err	t-test
ASC^2	constant for income level 2	-0.86	0.789	-1.09
ASC^3	constant for income level 3	-4.64	0.901	-5.14
ASC^4	constant for income level 4	-8.31	1.12	-7.39
ASC^5	constant for income level 5	-10.6	1.55	-6.82
β^3_{educ}	dummy for presence of people with higher educ 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
$\beta^2_{\text{zonal\_inc}}$	average zonal income	0.0008	0.0004	1.84
$\beta^3_{\text{zonal\_inc}}$	average zonal income	0.0012	0.0005	2.55
$\beta^4_{\text{zonal\_inc}}$	average zonal income	0.0016	0.0005	3.09
$\beta^5_{\text{zonal\_inc}}$	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
β^2_{workers}	number of workers in the household	1.14	0.277	4.11
β^3_{workers}	number of workers in the household	2.22	0.295	7.53
β^4_{workers}	number of workers in the household	2.46	0.345	7.13
β^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

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

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.