Simulation based Population Synthesis

Michel Berlaire, Bilal Farooq,
Ricardo Hurtubia, and Gunnar Flötteröd

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

February 22, 2013

Outline

Synthetic population

- Agent-based simulation
- Disaggregate models
- Need to access characteristics of agents

TRANSP-OR

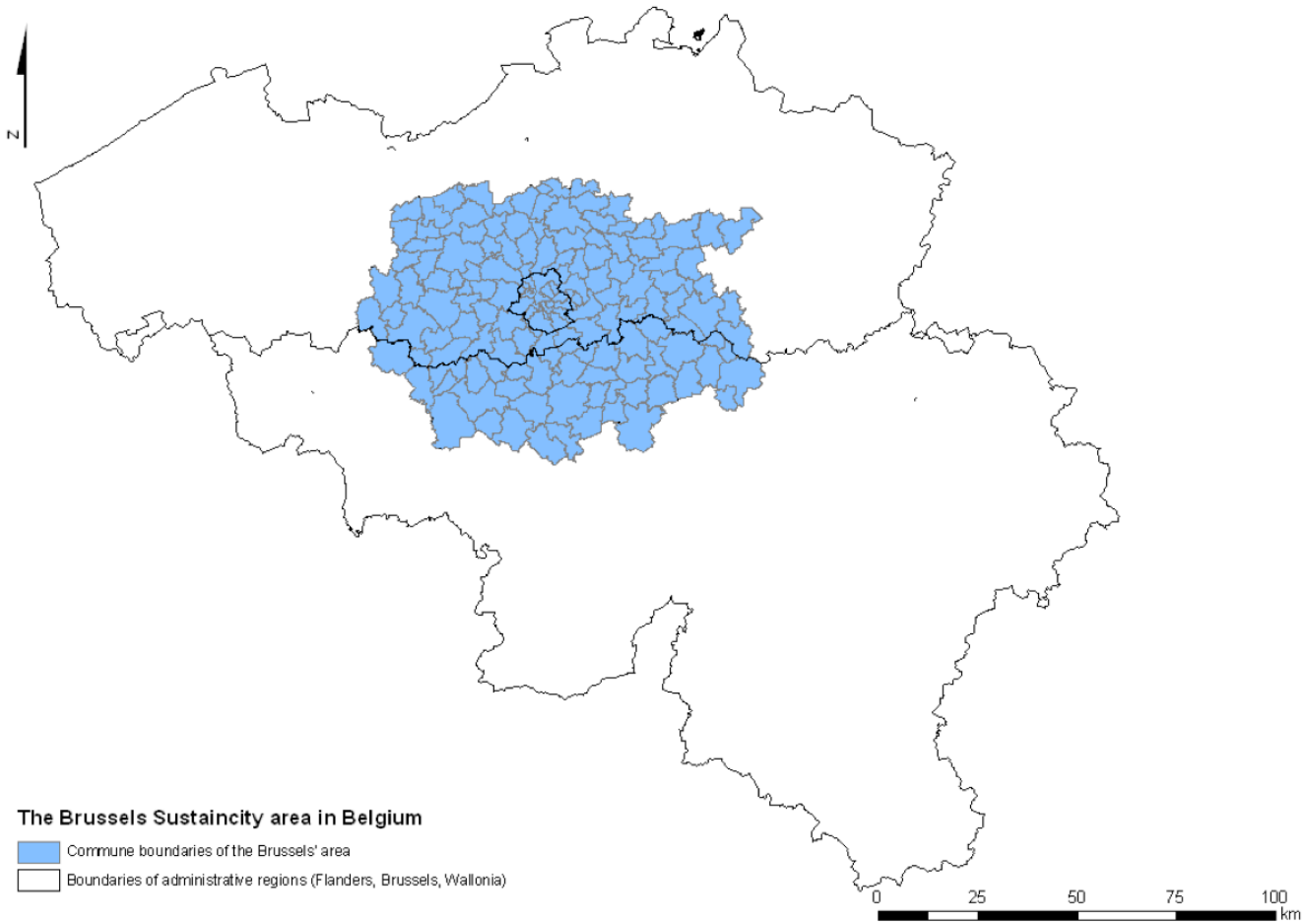
ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

SustainCity project

European Union project

- More than 10 major European universities involved
- Aims:
  - Integrated Land Use and Transportation modelling framework
  - Demographics, environment, and multi-scale issues
- Tools: Urbansim, MATSIm, Metropolis
- Case studies
  - Paris
  - Zurich
  - Brussels

SustainCity: Brussels case study



TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

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 characteristics

- Size, children, workers, cars, income, university education, dwelling type, sector

- *Conventional synthesis procedures were not usable*

Agent synthesis in transportation

From *Four-Stage* to *Activity based Integrated* modelling

- Urban Microsimulation

Forecasting behavior using individual level models

- Lack of individual level data for population
- Synthesis of individual agents and their characteristics

Initial work

- TRansportation ANalysis SIMulation System (TRANSIMS) [LaRon et al., 1996]
- Focused on synthesis of a small sub-set of characteristics

Current needs and trends (MATSim, ILUTE, etc.)

More detailed characteristics

- e.g. social network, dwelling location, employment type

Associations between different types of agents

- Household, family, person

Using variety of different data sources at different spatial aggregation

- From samples to aggregate statistics
- Sources: Census, travel survey, household spending survey, labor force survey, statistics from revenue agency, real estate cadaster, ...
- Space: traffic analysis zone, census sector, commune, municipality, ...

Existing approach

- **Fitting based approach**
  - Iterative proportional fitting
  - 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

- Sample used to initialize the contingency table
- Use marginal as dimensional totals
- Adjust the cell probabilities 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

Iterative Proportional Fitting

Most widely used in transportation literature

Historic improvements in IPF

Zero-cell issue, dimensionality, fractions in CT, zone-by-zone vs entire area, associations, sample-less, ...

Literature

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), Barthelemy and Toint (2012), Müller and Axhausen (2012) HIPF, ...

Combinatorial Optimization (CO) [Williamson, 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

- Amplification of errors
- Lack of heterogeneity

Focus on fitting marginals

- No emphasis on the correlation structure

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 characteristics only
- Scalability issues

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Problem statement

True population

- Individual agents defined as a set of characteristics $X = (X^1, X^2, ..., X^n)$
- Discrete (e.g. marital status) or continuous (e.g. income)
- Unique joint distribution represented by $\pi_X(x)$

Complex distribution

- No direct access to $\pi_X(x)$
- Hard to draw from

Partial views of $\pi_X(x)$

- Marginals, conditional-marginals, and samples

Problem statement

Develop a synthesis procedure that...

- draws a synthetic population
- uses the partial views as if we were drawing from $\pi_X(x)$
- generates an empirical distribution $\pi_{\widehat{X}}(\widehat{x})$ of $\widehat{X}$ close to $\pi_X(x)$

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Simulation based approach

Gibbs sampling

- Markov Chain Monte Carlo [Geman & Geman, 1984]
- draws from complex distributions: $\pi_X(x)$
- exploits conditionals

$$\pi(X^i|X^j = x^j, \text{ for } j = 1...n \ \& \ i \neq j) = \pi(X^i|X^{-i}), \text{ for } i = 1, ..., n$$

Key challenges

- Preparation of the conditional distributions for characteristics from available data sources
- Full-conditionals rarely available

Completing conditionals by assumptions

Example: (Age | Sex, Income)

- From data only (Age | Income) available
- Assume that for all values of Sex:

$$(\text{Age} \mid \text{Sex}, \text{Income}) = (\text{Age} \mid \text{Income})$$

- No matter the Sex of a person, Age is only dependent on Income

Completing conditionals by assumptions

Required

$$\pi(X^1|X^{-1}) = \pi(X^1|X^{(2...k)}, X^{((k+1)...n)})$$

Available

$$\pi(X^1|X^{(2...k)})$$

Assumption

$$\pi(X^1|X^{-1}) = \pi(X^1|X^{(2...k)}), \forall X^{((k+1)...n)}$$

Worst case

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

Completing conditionals by domain knowledge

Example: (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 domain knowledge

Expert's assumptions

$$\pi(X^1|X^{(2\ldots k)}, X^{((k+1)\ldots n)} = a) = \pi^a(X^1|X^{(2\ldots k)}),$$
$$\pi(X^1|X^{(2\ldots k)}, X^{((k+1)\ldots n)} = b) = \pi^b(X^1|X^{(2\ldots k)}),$$
$$\ldots$$

Completing conditionals by parametric models

Example: (Dwelling | Income , Sex, Age)

- In sample
  - Characteristics of person n (Dwelling , Age, Sex)$_n$
  - Attributes of zone (z)
- Dwelling choice model can be estimated
  - Choice set: $dwel\_typ = (attached, semidetached, detached, apartment)$
  - Utility:
    $V^i_{(n,z)} = ASC^i + \beta^i_{age_n} \times Age + \beta^i_{av\_inc_z} \times av\_inc_z + interactions + ...$
  - Model: logit.

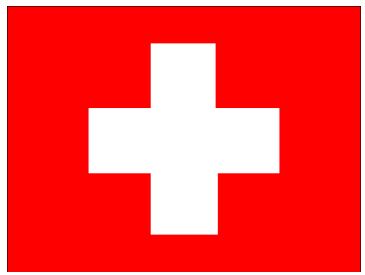
Completing conditionals by parametric models

Logit model

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

Population from Swiss Census

- Access to Swiss Census for 2000
  - Person and household characteristics (Except for Income)
- Selected area: postal code in Lausanne
  - CH-1004
  - 28,533 persons
- Four person characteristics (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)



TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

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

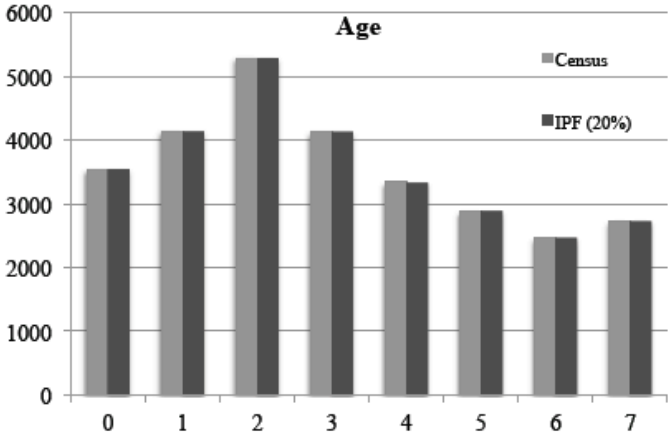
List of available conditionals

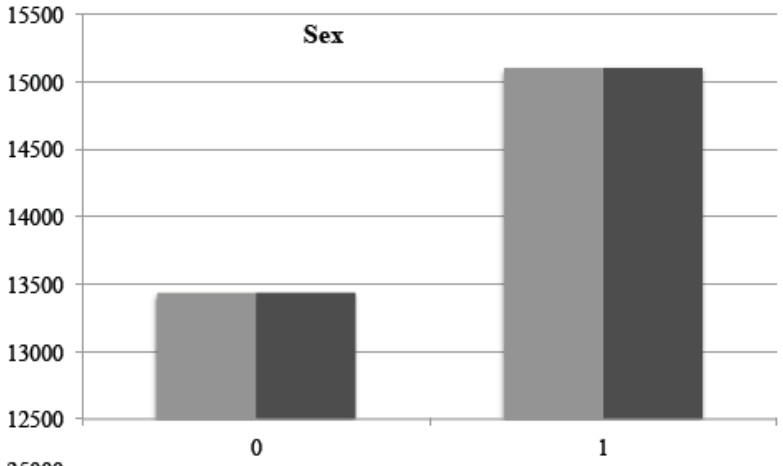
No.	ID	Conditionals
1	*FullCond*	$\pi(age \mid sex, hhld\_size, edu\_level)$ $\pi(sex \mid age, hhld\_size, edu\_level)$ $\pi(hhld\_size \mid age, sex, edu\_level)$ $\pi(edu\_level \mid age, sex, hhld\_size)$
2	*Partial_1*	$\pi(age \mid \cancel{sex}, hhld\_size, edu\_level)$ $\pi(sex \mid age, hhld\_size,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

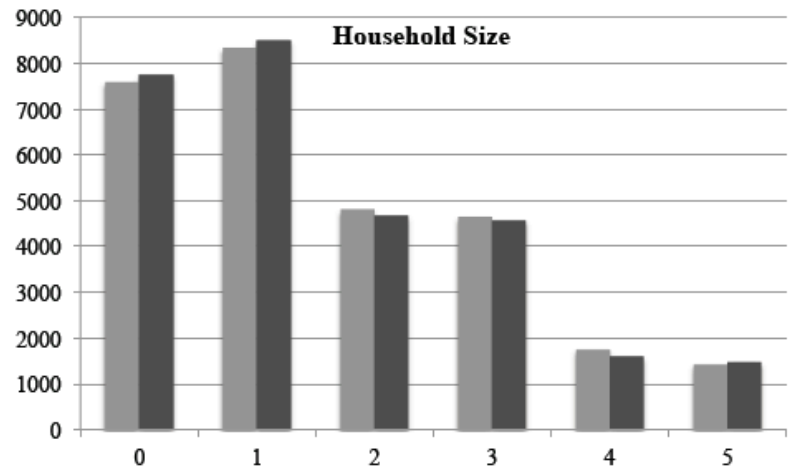
Data preparation

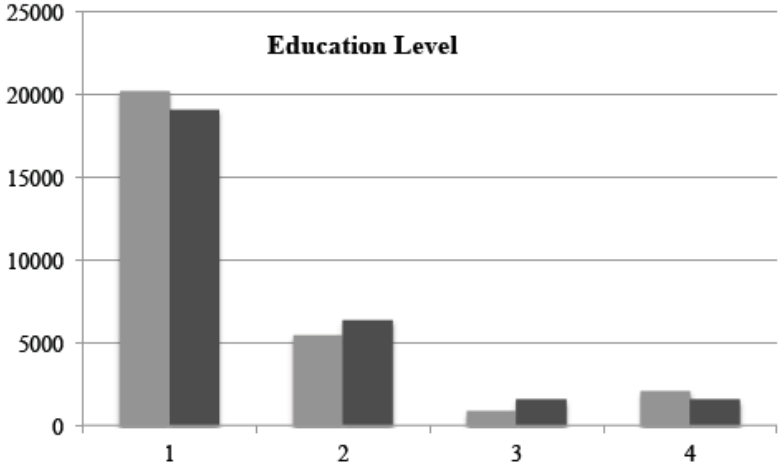
- Based on sample-conditional combinations
  - 20 possibilities
- IPF can use marginals only
  - Number of experiments collapses to 5
- Simulation based synthesis
  - Used conditionals only (used lesser information)
  - Number of experiments collapses to 4

Results: IPF and Census marginals

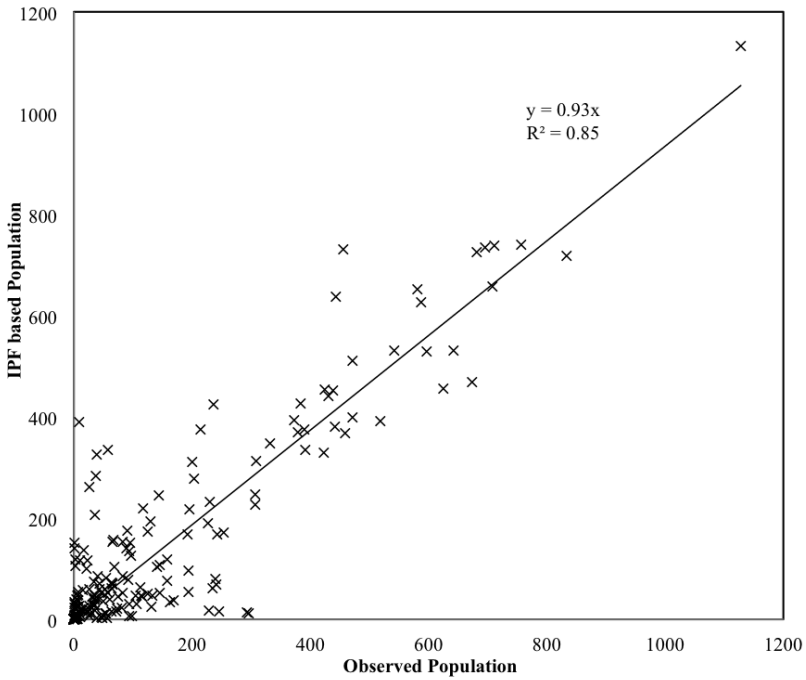




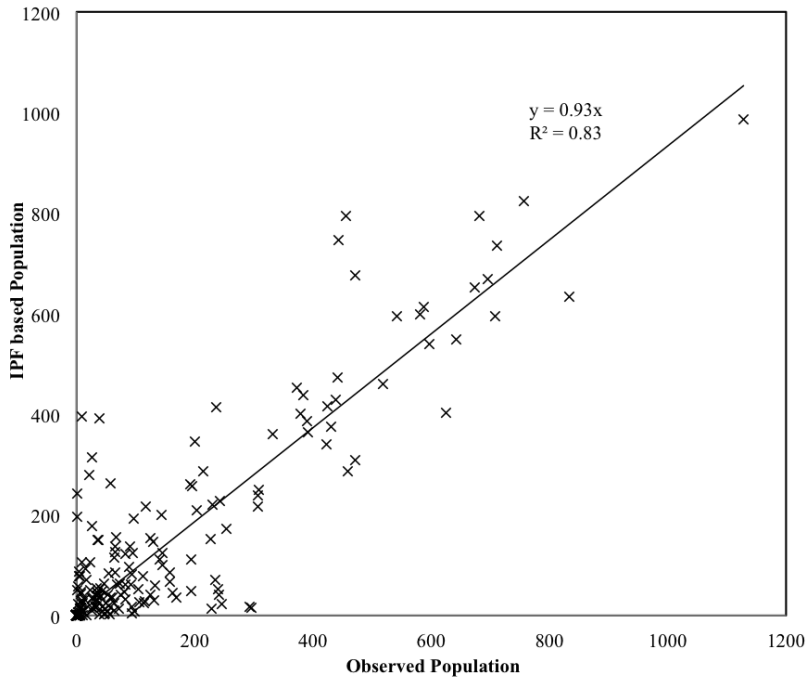




Results: Fit of IPF with Census joint distribution

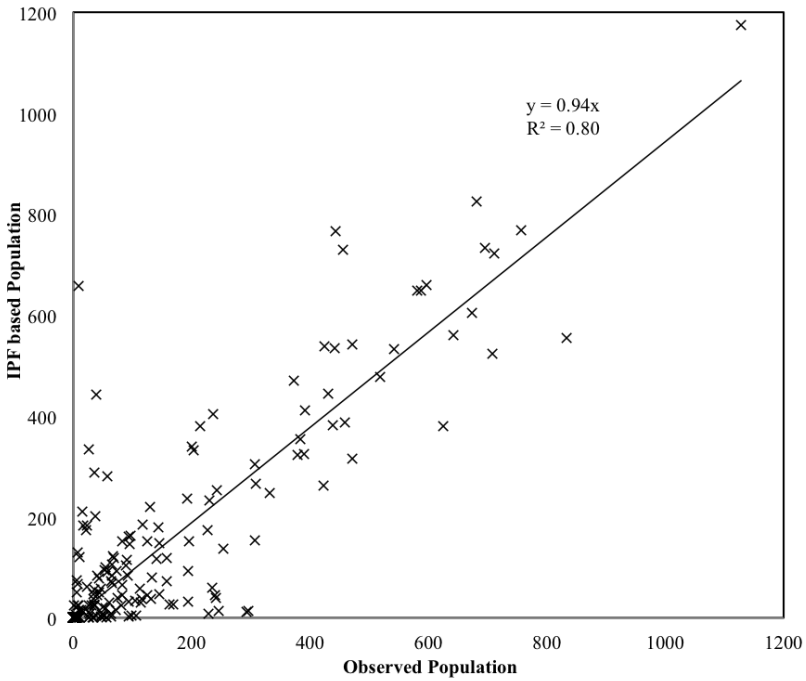


IPF with 20% sample

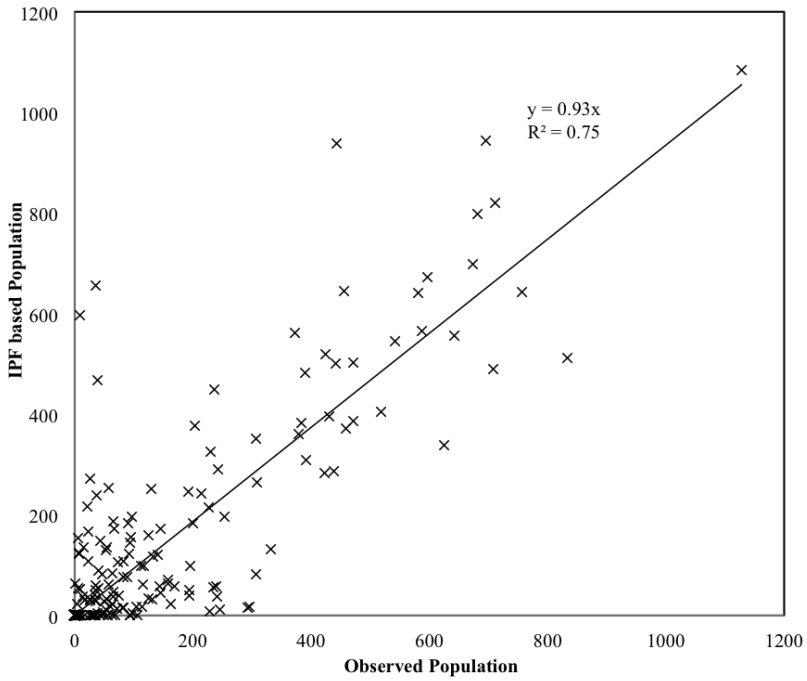


IPF with 10% sample

Results: Fit of IPF with Census joint distribution

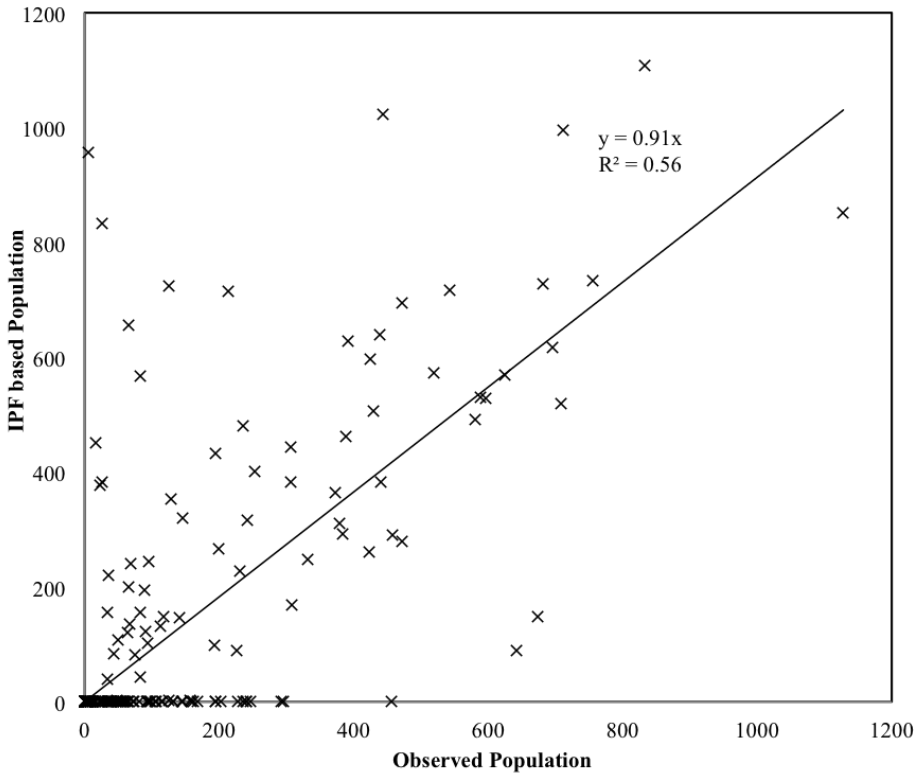


IPF with 5% sample



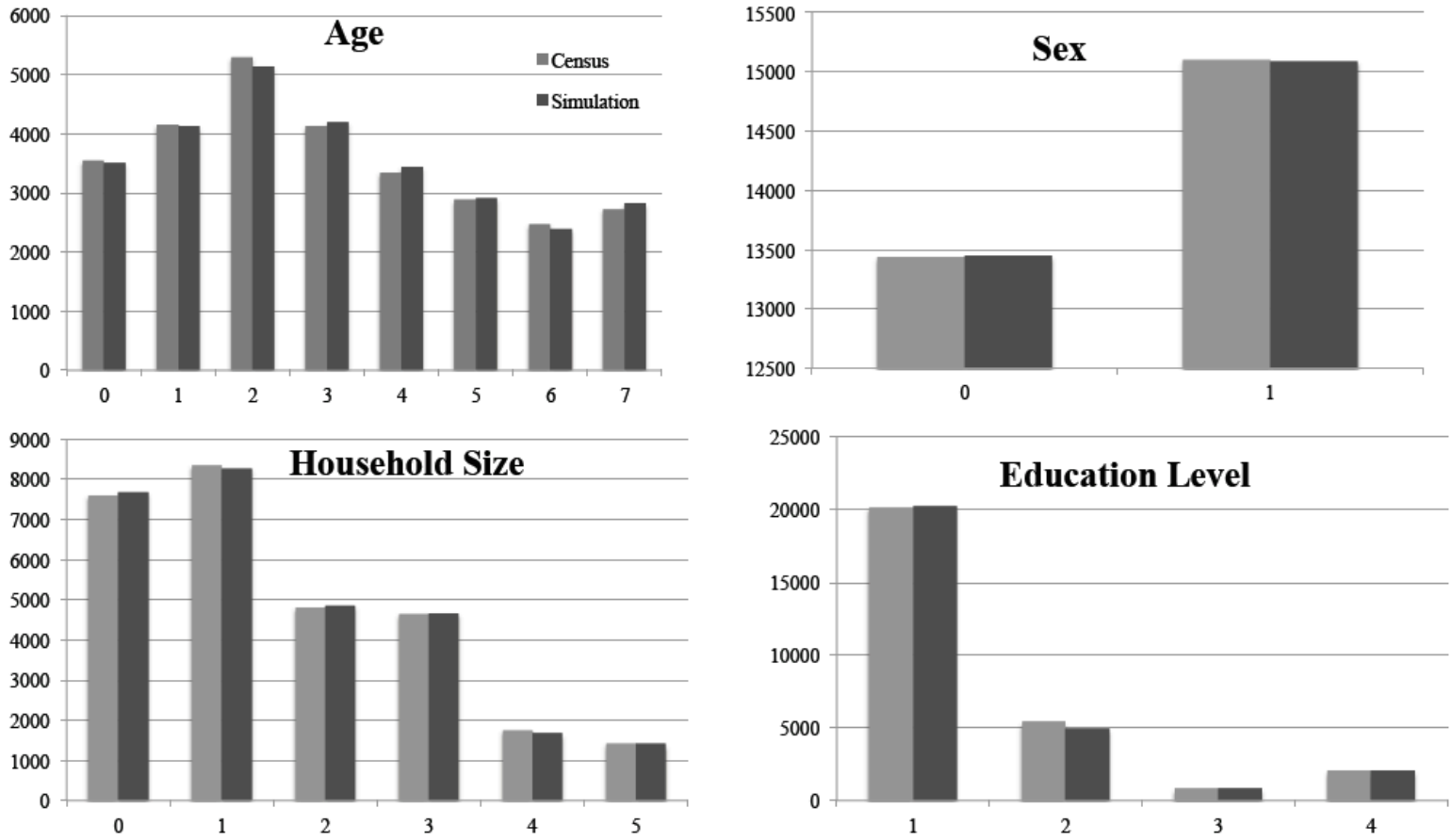
IPF with 3% sample

Results: Fit of IPF with Census joint distribution



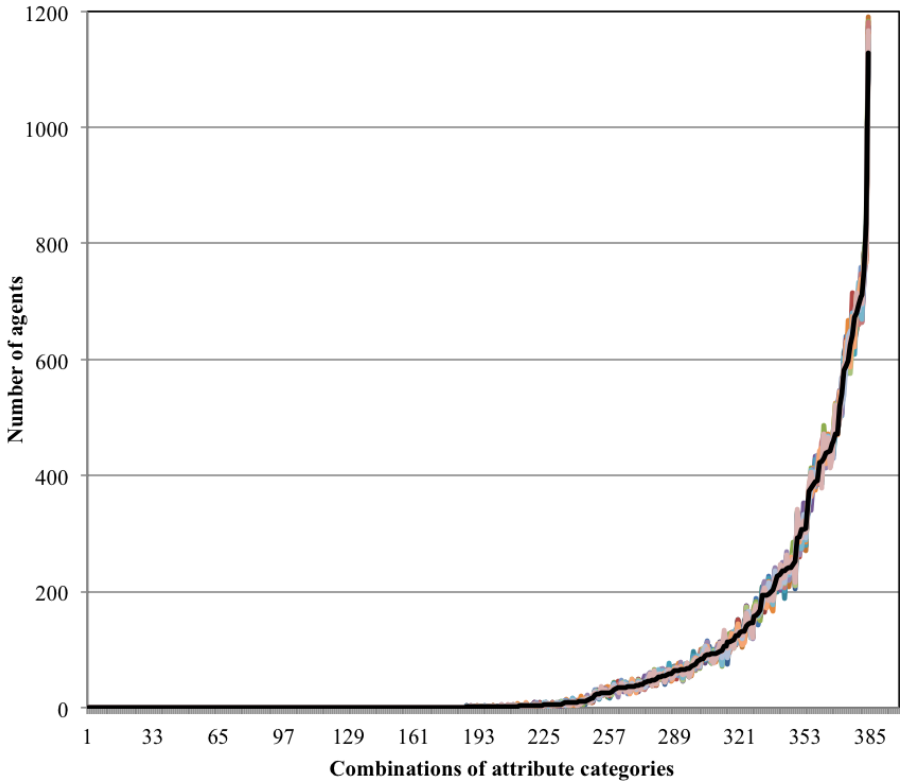
IPF with 1% sample

Results: Simulation and Census marginals



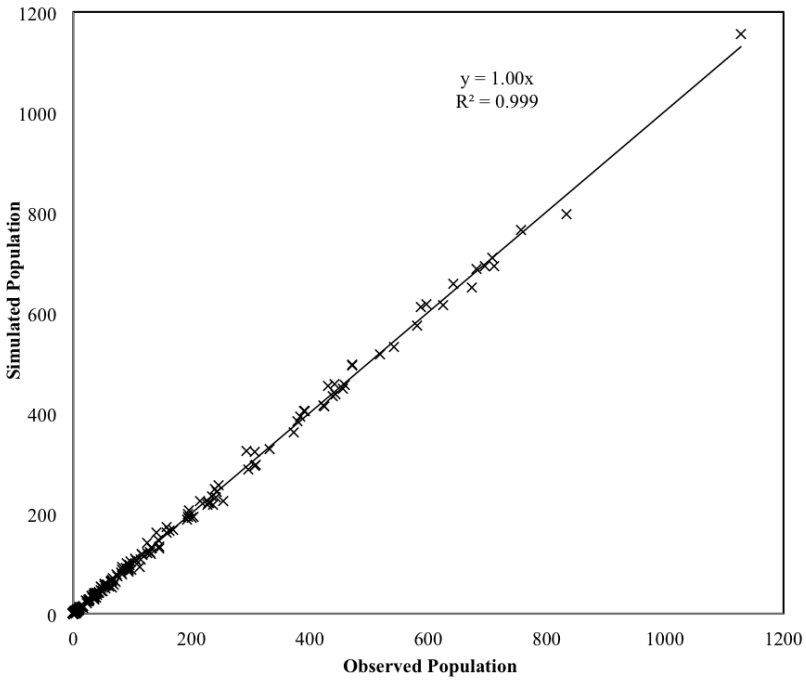
Using full-conditionals (*FullCond*)

Results: Simulation and Census joint dist.

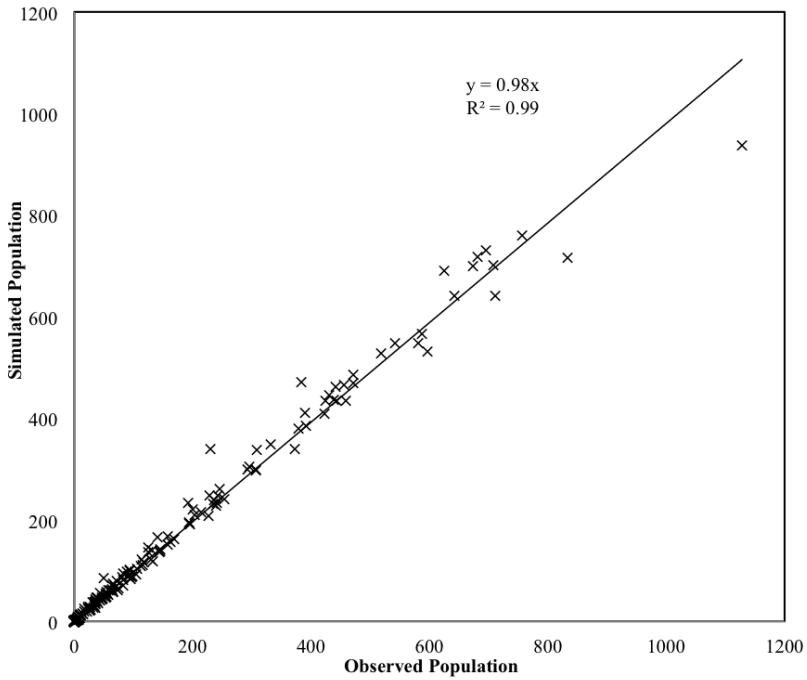


20 runs based on *FullCond* with real population superimposed

Results: Fit of Simulation with Census joint dist.



FullCond

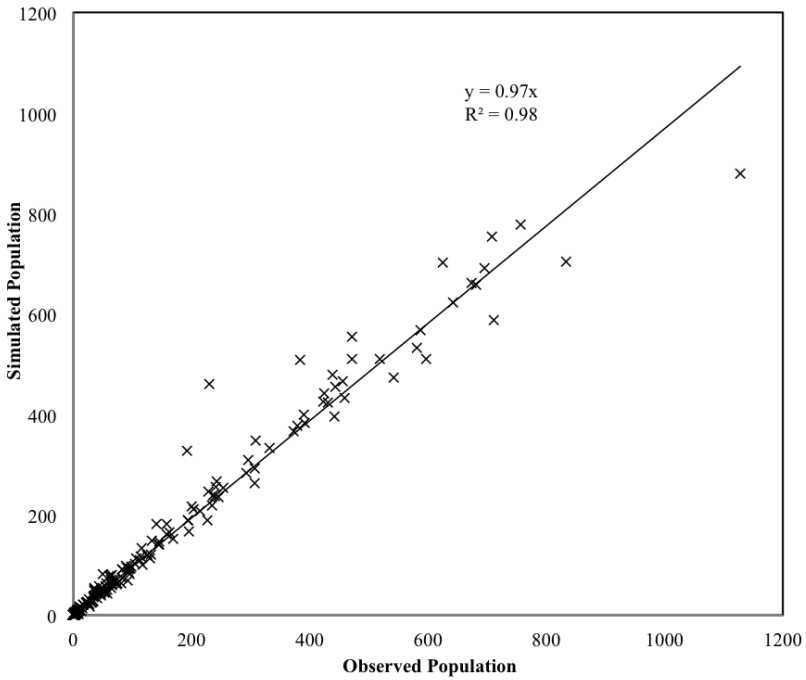


Partial_1 (Sex missing in 1 conditional)

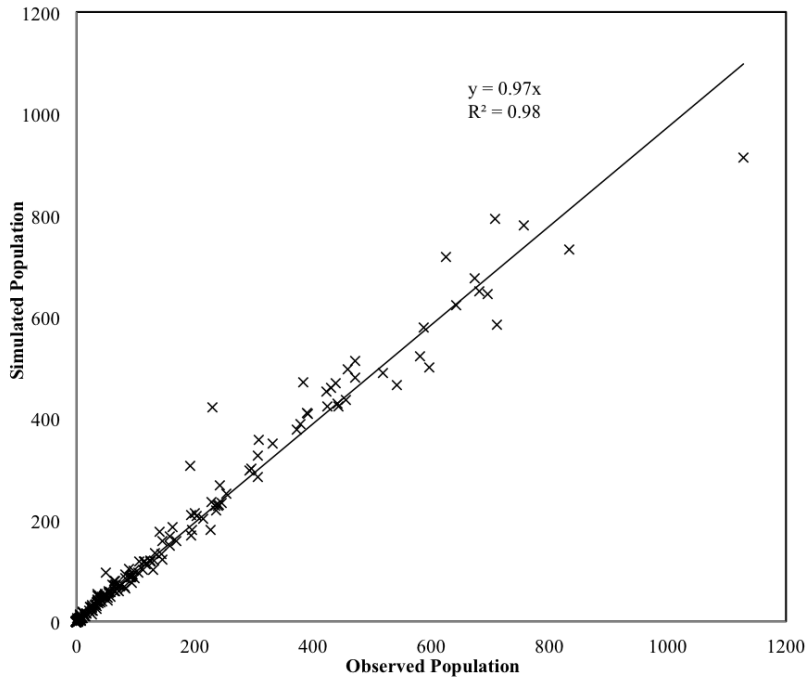
TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

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

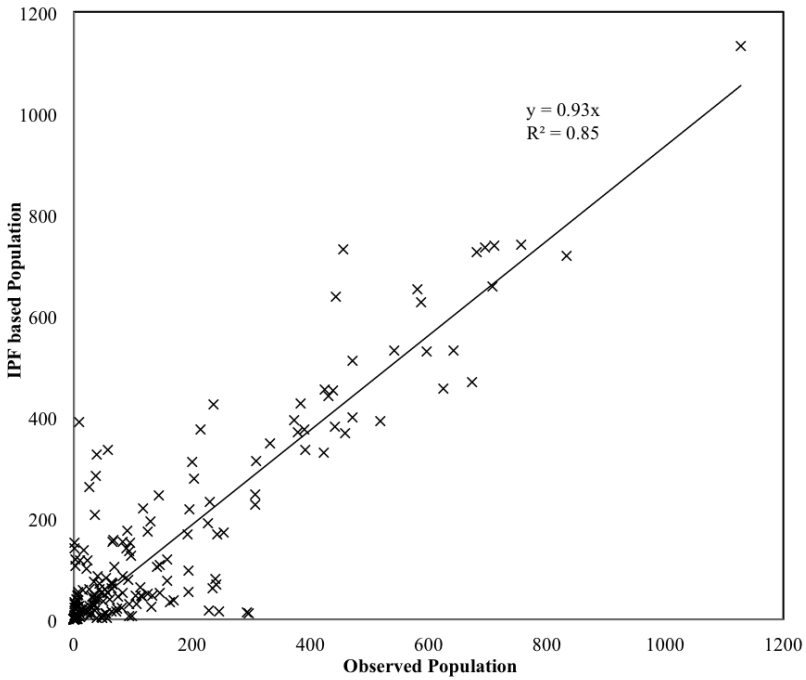
Comparison: Standard Root Mean Square Error

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

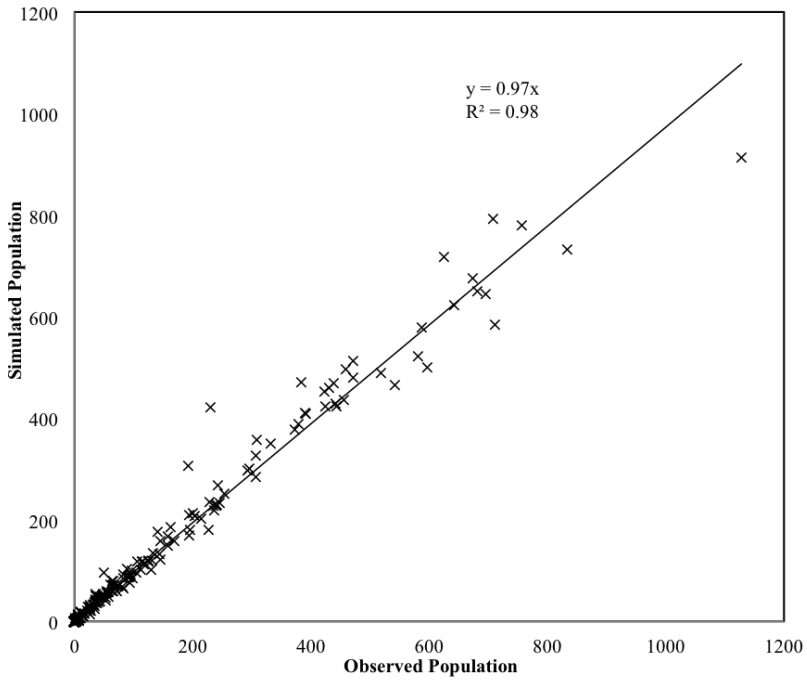
Input	IPF	Simulation
20%*Sample*	0.853	-
10%*Sample*	0.928	-
5%*Sample*	1.020	-
3%*Sample*	1.160	-
1%*Sample*	1.730	-
FullCond	-	0.130
Partial_1	-	0.240
Partial_2	-	0.340
Partial_3	-	0.350

- For Marginals only, both methods give the same fit

Best case IPF and worst case Simulation

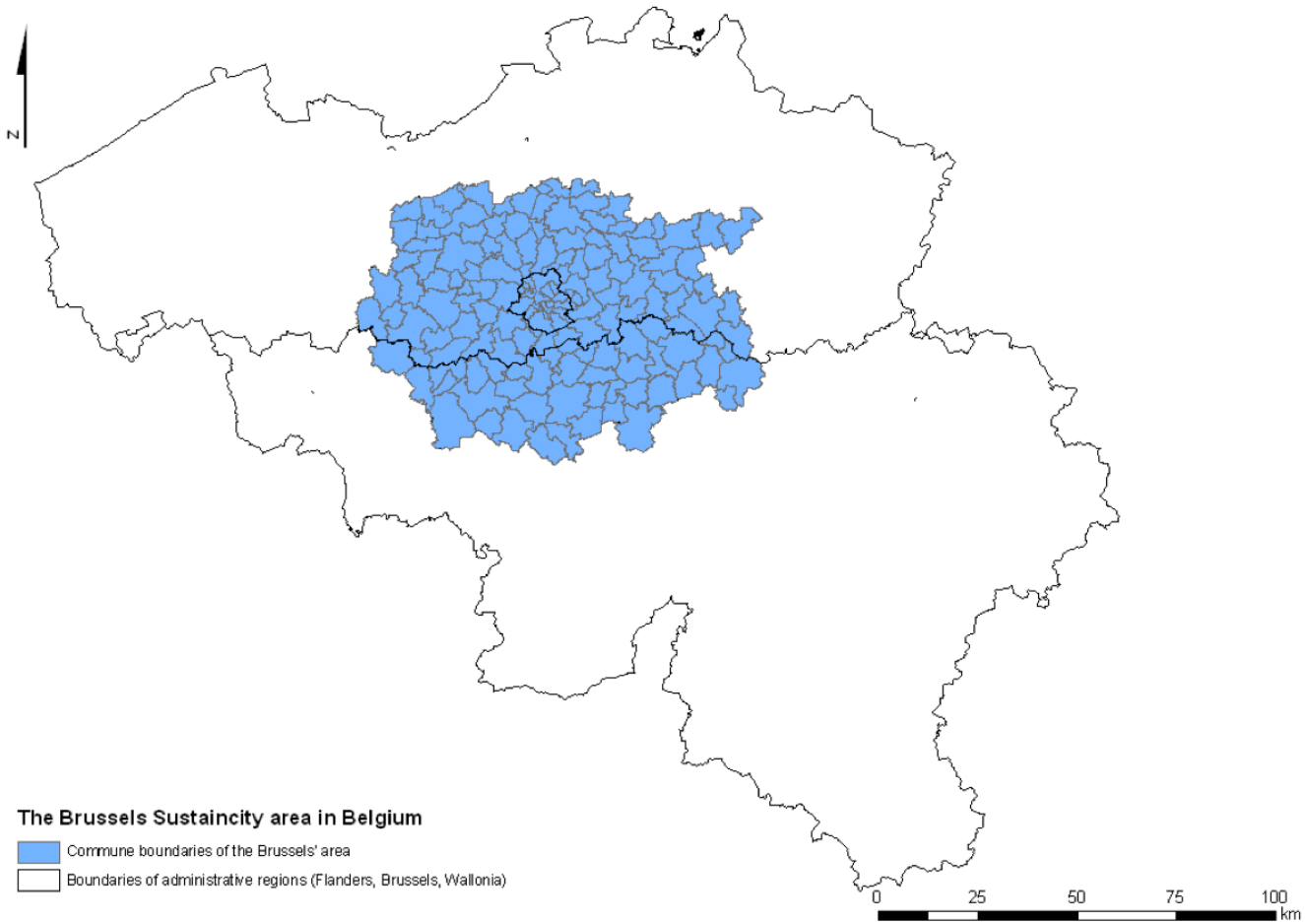


IPF with 20% sample



Partial_3 (Sex missing from all the conditionals)

Back to Brussels case study



TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

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)

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

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

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

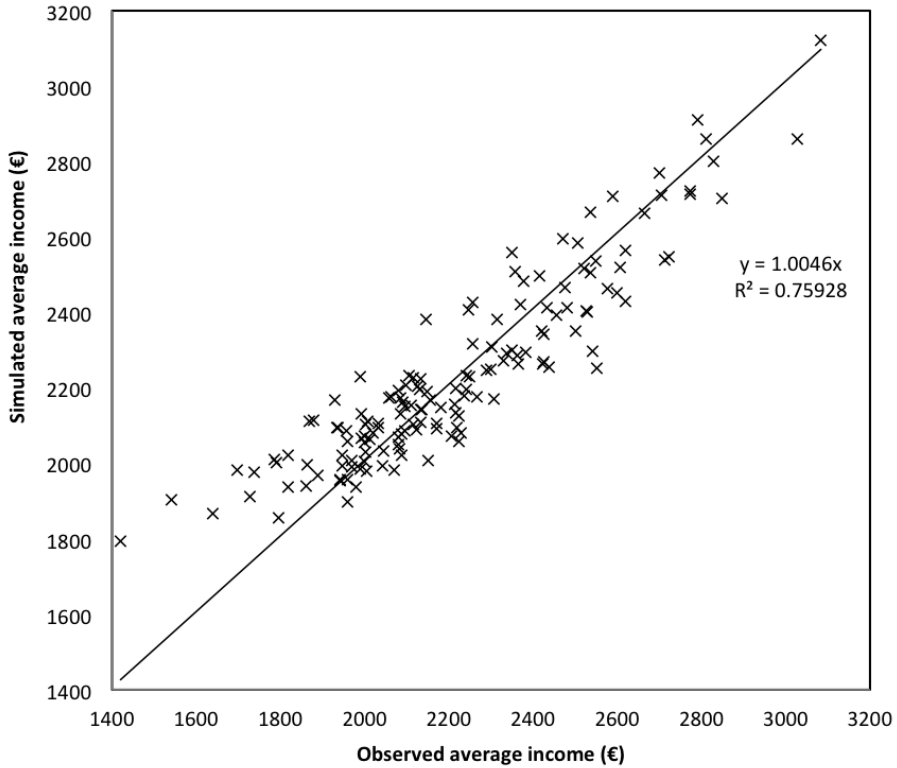
$$V^4_{(hh,z)} = ASC^4 + \beta^4_{educ_{hh}} \times educ_{hh} + \beta^4_{zonal\_inc_z} \times zonal\_inc_z + \beta^4_{cars_{hh}} \times cars_{hh} + \beta^4_{house_{hh}} \times house_{hh} + \beta^4_{workers_{hh}} \times workers_{hh}$$

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

Income level model

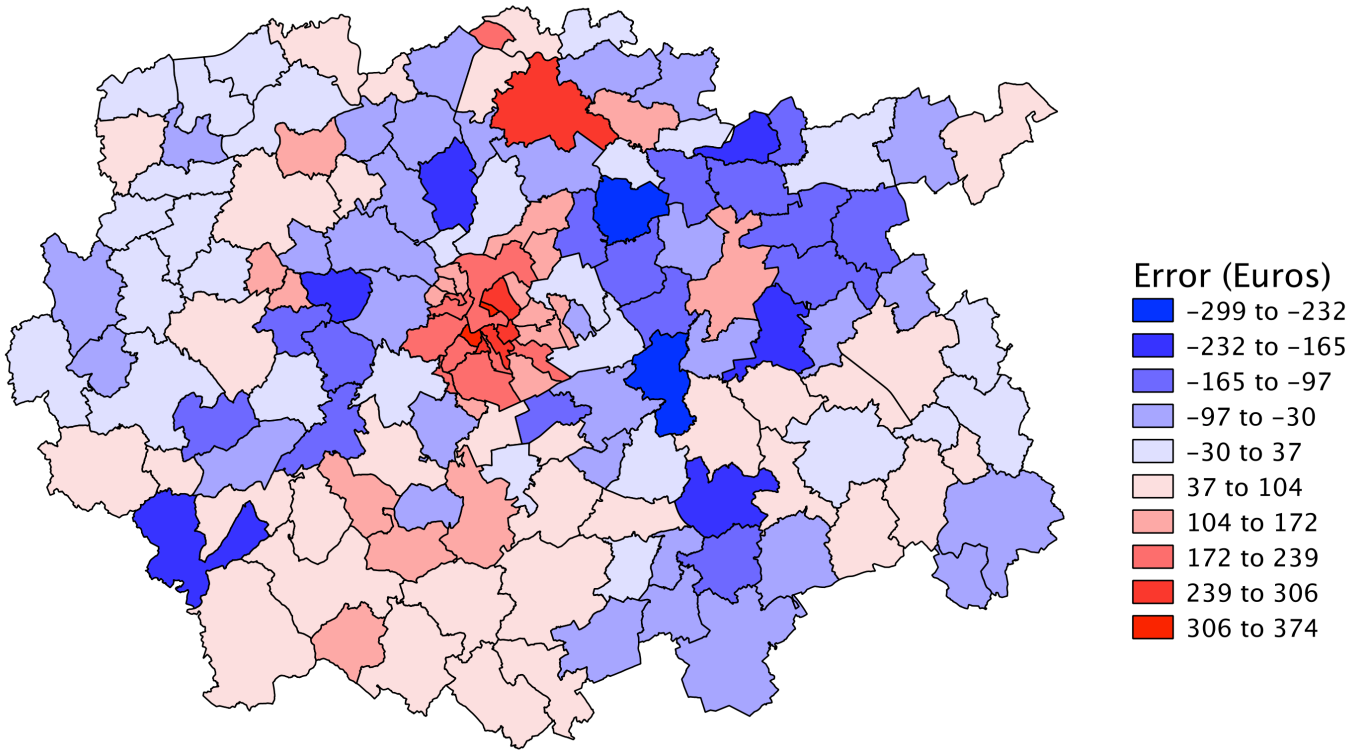
Parameter	Variable	Value	Std err	t-test
ASC^2	constant for income level 2	-0.86	0.789	-1.09
ASC^3	constant for income level 3	-4.64	0.901	-5.14
ASC^4	constant for income level 4	-8.31	1.12	-7.39
ASC^5	constant for income level 5	-10.6	1.55	-6.82
β^3_{educ}	dummy for presence of people with higher educ 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
$\beta^2_{\text{zonal\_inc}}$	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

Concluding remarks

- From single solution optimization problem to sampling from joint distribution
  - Output of Land Use and Transport models

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

- 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

Concluding remarks

- Separation of data preparation from agent generation
  - Data, models, assumptions
- 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
- Generic data fusion technique
- Technical report URL: `http://transp-or.epfl.ch/documents/technicalReports/FaroBierHurtFloe_PopSyn2013.pdf`

Acknowledgments

- This research is funded by
  - European Commission's Seventh Framework Programme
  - Swiss National Science Foundation
  - Danish Council of Strategic Research
- Many thanks to
  - *Tomáš Robenek* for help in developing Java version
  - *Sohrab Sahaleh* for data processing
  - *Lovisa Arnesson* for data processing and running experiments