Life-Course Synthetic Populations

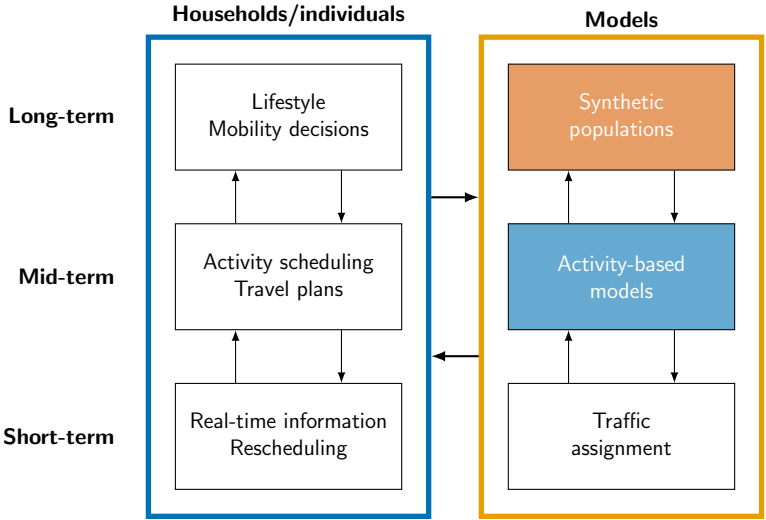
Candice Baud Michel Bierlaire

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

June 24, 2026

Outline

Travel demand modeling



Outline

Synthetic populations

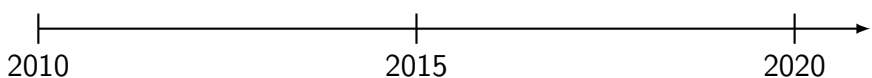
Cross-sectional

- Snapshot of the population at a given point in time.
- Based on an observed real population (census).
- Share the same statistical properties as the real population.
- Includes the status of long-term mobility decisions: home and work location, vehicle ownership, driver's license ownership, etc.
- Feed into activity scheduling models.

Multiperiod synthetic populations

Challenges

- Lack of panel data.
- Instead, repeated cross-sectional census data.
- Consistency (not necessarily the same individuals).



Traditional synthetic populations

Static

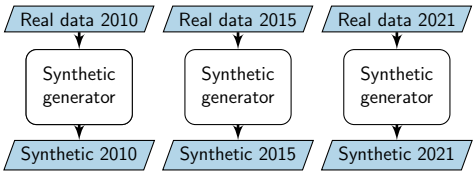
- Sex
- Age
- Income
- Employment status
- Level of education
- Home location
- Work location
- "Mobility tools" ownership
- Driver license
- etc.

Dynamic

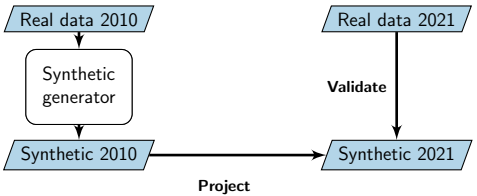
- Sex
- Age(t)
- Income(t)
- Employment status(t)
- Level of education(t)
- Home location(t)
- Work location(t)
- "Mobility tools" ownership(t)
- Driver license(t)
- etc.

Traditional synthetic populations

Static



Dynamic



Traditional synthetic populations

Static

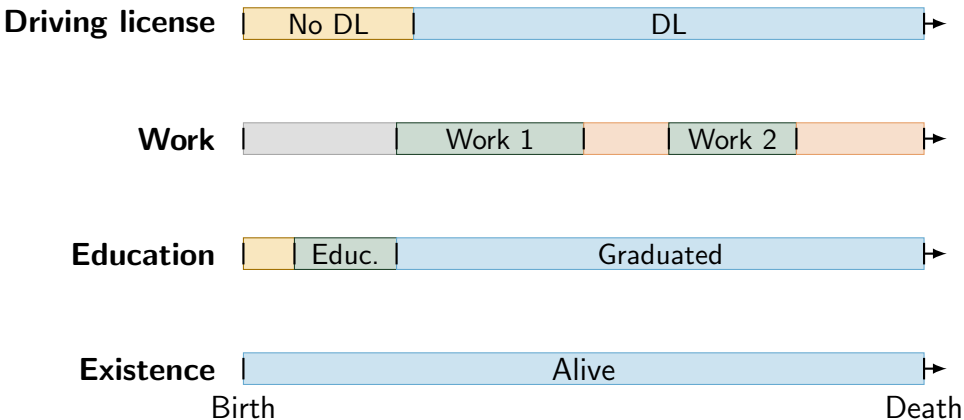
- Iterative Proportional Fitting.
  [Beckman et al., 1996]
- Combinatorial Optimization.
  [Abraham et al., 2012]
- Simulation-based. [Farooq et al., 2013]
- Machine Learning.
  [Xu and Veeramachaneni, 2018]

Dynamic

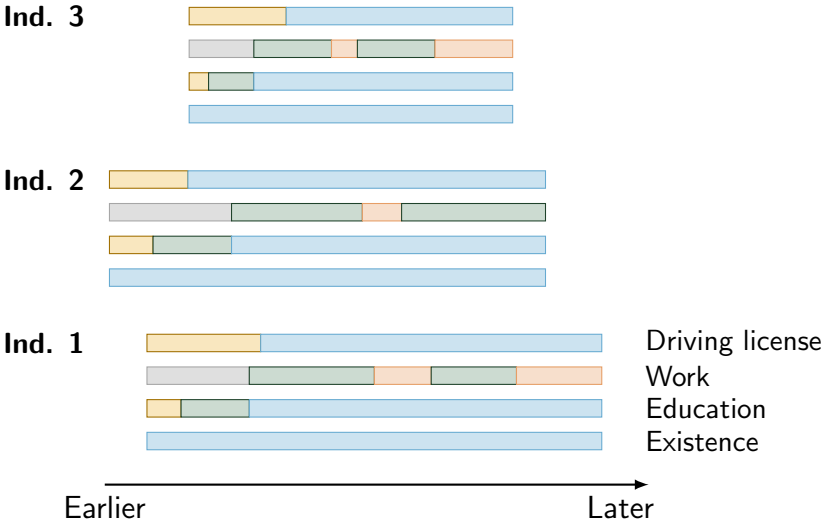
- Dynamic projection.
  [Namazi-Rad et al., 2014]
- Static projection. [Lomax et al., 2022]
- Resampling.
  [Prédhumeau and Manley, 2023]
- Hybrid approaches. [Kukic et al., 2023]

Outline

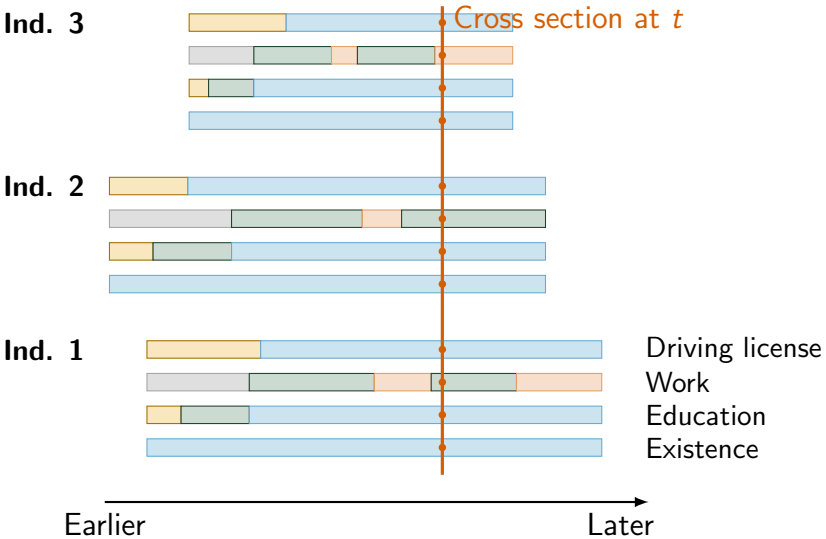
Life as parallel temporal streams



A synthetic population as a collection of life trajectories



A cross section of the synthetic population at time t



Life-course representation

Individual life course

$$X_n = \left(X_n^{\text{edu}}, X_n^{\text{work}}, X_n^{\text{res}}, X_n^{\text{lic}}, \ldots\right)$$

$$Y_n(t) = T(X_n, t)$$

- X_n: complete life-course description, independent from t.
- $Y_n(t)$: state observed at time t.

Example

$$X_n = \begin{pmatrix} \text{date of birth} \\ \text{lifespan} \\ \text{education history} \\ \text{work history} \\ \text{license history} \\ \vdots \end{pmatrix}$$

$$\text{Age}_n(t) = t - \text{date of birth}_n$$

Feasible life courses

Structural constraints

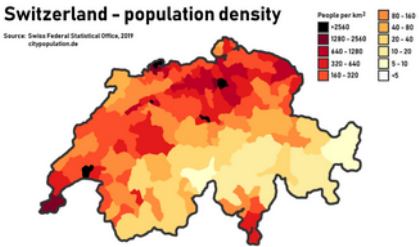
- Each dimension is a sequence of events.
- Events cannot overlap on the same dimension.
- Durations must satisfy feasibility constraints.
- The sequence must be temporally consistent.

Dimension-specific state

$$Y_n^k(t) = T_k(X_n^k, t)$$

Outline

Data

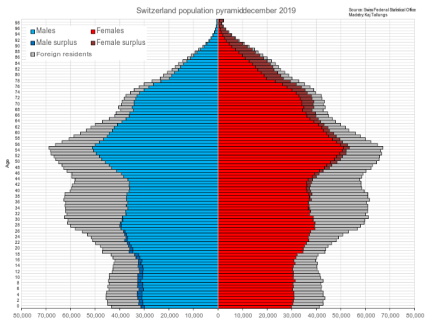


!

Census data

- Not necessarily available every year.
- Not necessarily exhaustive.
- Usually cross-sectional.
- Switzerland: every 5 years, $\approx 5\%$ sample.

Prior models for life courses



Role of the prior

- Priors describe plausible life courses before observing the data.
- They are defined on the time-independent variables X.
- They combine demographic models, behavioral models, and duration models.

Prior models for life courses

Examples

- Age(t): birth date and lifespan [Gompertz, 1833].
- Income(t): income evolution models [Kaldasch, 2012].
- Employment status(t): choice of employment status [Kolvereid, 1996].
- Level of education(t): educational choice models [Manzo, 2013].
- Home location(t): last location, moving behavior [de Palma et al., 2015].
- Work location(t): firm relocation [Bodenmann and Axhausen, 2015].
- "Mobility tools" ownership(t): last vehicle, duration model [Gilbert, 1992].
- Driver license(t): date of acquisition [Nurul Habib, 2018].

Example: lifespan prior

[Gompertz, 1833] distribution

For the lifespan L, one possible prior is

$$\Pr(L \leqslant \ell) = 1 - \exp\left(-b\,\frac{\exp(\eta\ell) - 1}{\eta}\right), \qquad \ell \geqslant 0.$$

- $b > 0$: scale parameter.
- $\eta > 0$: shape parameter.

Why priors are not enough

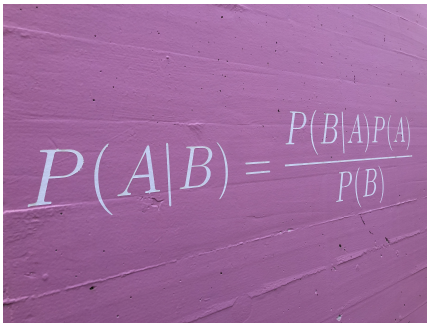
Limitations

- Prior distributions may not reflect the true population.
- They may be outdated.
- Structural changes may occur.
- Examples include pandemics, policy changes, and long-term demographic shifts.

Consequence

Observed cross-sectional data are used to update the prior distribution of life courses.

Bayesian formulation



Bayes theorem

- X: complete life-course synthetic population.
- B: observed cross-sectional data.

Posterior distribution

$$f(X \mid B) \propto \mathcal{L}(B \mid X)\, f_{\text{prior}}(X)$$

- $f_{\text{prior}}(X)$: prior plausibility of the life courses.
- $\mathcal{L}(B \mid X)$: agreement with observed data.

Likelihood from cross-sectional data

Projection

- A life-course population X is projected to observation time t:

$$Y_t = T(X, t).$$

- The observed data B_t are treated as noisy observations of Y_t:

$$B_t = Y_t + \varepsilon.$$

Likelihood contribution

For an observed quantity B_{tk},

$$P(B_{tk} \mid X) = f_\varepsilon \left(T_k(X, t) - B_{tk} \right).$$

Population size and identifiability

Issue

- The likelihood only involves individuals alive at time t.
- Non-alive individuals do not affect $\mathcal{L}(B_t \mid X)$.
- The posterior is weakly identified.

Regularization

- Introduce a target population size over time.
- Add a penalty term:

$$-\frac{1}{2\sigma^2} \sum_k w_k \left(\tilde{M}_{t_k}(X) - M^\star_{t_k} \right)^2$$

- This stabilizes the estimation.

Sampling strategy

Posterior sampling

- The objective is to draw from

$$f(X \mid B) \propto \mathcal{L}(B \mid X)\, f_{\text{prior}}(X).$$

- Different components of X require different proposal mechanisms.

Sampling scheme

- Event occurrence indicators: Gibbs sampling.
- Event times and durations: Metropolis–Hastings.
- Constrained continuous variables: hit-and-run Metropolis–Hastings.

Hit-and-run Metropolis–Hastings

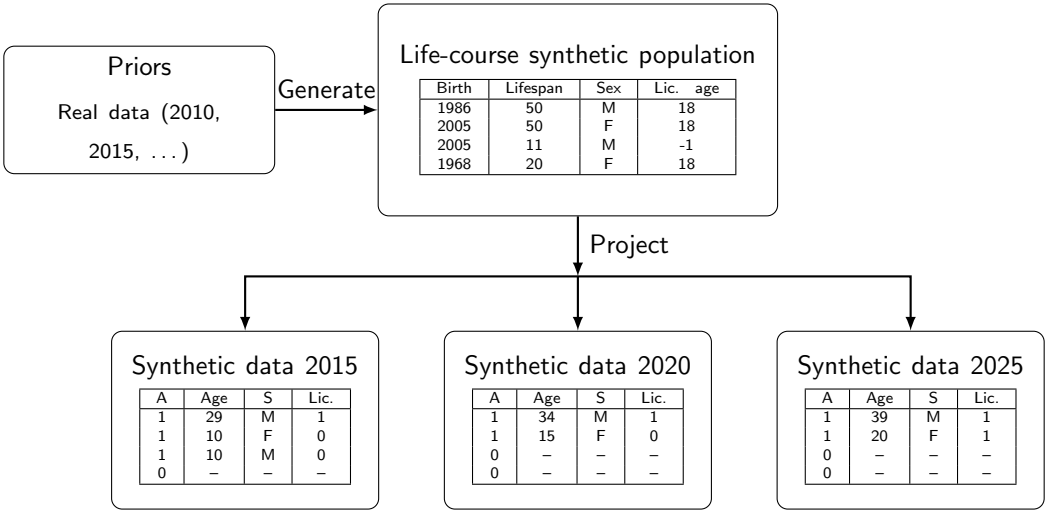
Feasible set

- Continuous variables must satisfy linear constraints.
- The feasible region is a convex polytope.
- Examples:
  - no driving license before age 18;
  - synchronized residence and employment spells.

Proposal

- Draw a random direction.
- Compute the feasible segment.
- Draw a candidate uniformly on the segment.
- Accept or reject with a Metropolis–Hastings step.

From life courses to synthetic cross-sections



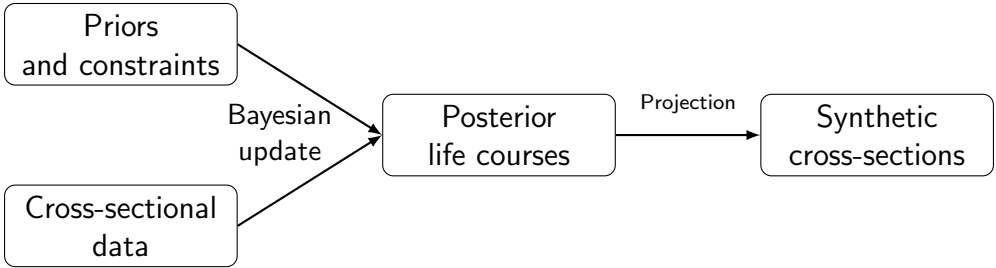
Birth	Lifespan	Sex	Lic. age
1986	50	M	18
2005	50	F	18
2005	11	M	-1
1968	20	F	18

A	Age	S	Lic.
1	29	M	1
1	10	F	0
1	10	M	0
0	–	–	–

A	Age	S	Lic.
1	34	M	1
1	15	F	0
0	–	–	–
0	–	–	–

A	Age	S	Lic.
1	39	M	1
1	20	F	1
0	–	–	–
0	–	–	–

Method summary



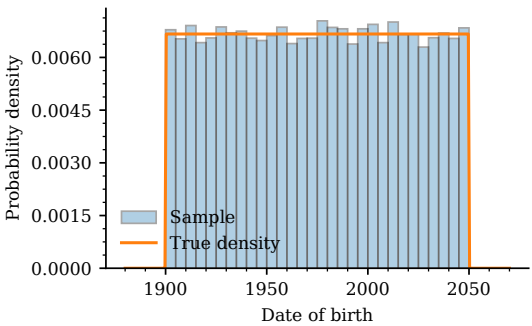
$$f(X \mid B) \propto \mathcal{L}(B \mid X)\, f_{\text{prior}}(X)$$

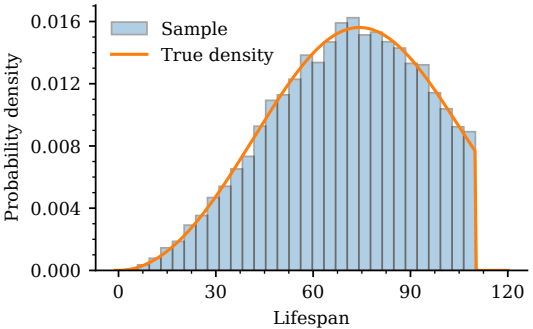
Main idea

- Generate complete life courses.
- Update them with cross-sectional observations.
- Project them to any year of interest.

Outline

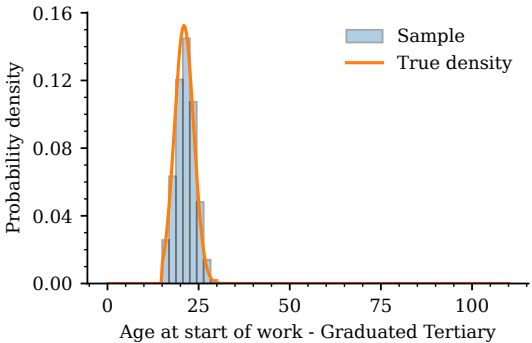
Priors: Existence

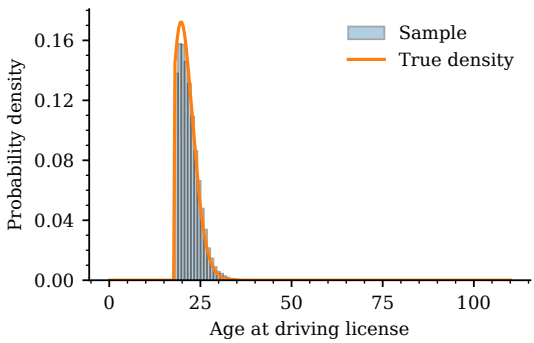




Date of birth and Lifespan sampled and true densities

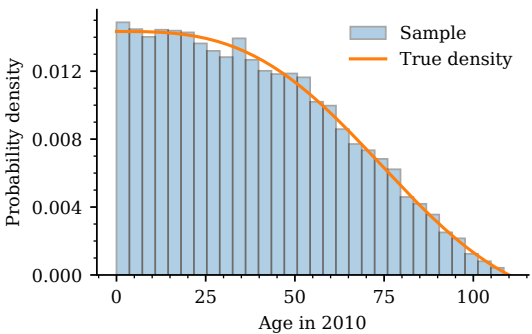
Priors: Other dimensions

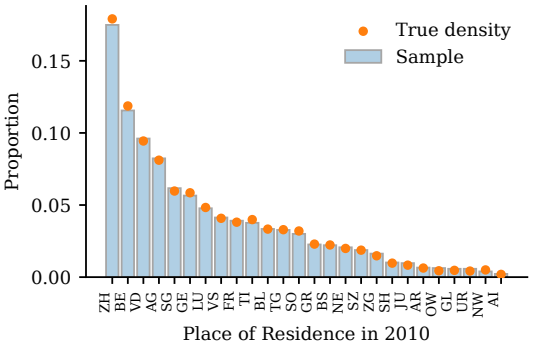




Age at labor entry (Graduated from Tertiary) and Driving license age

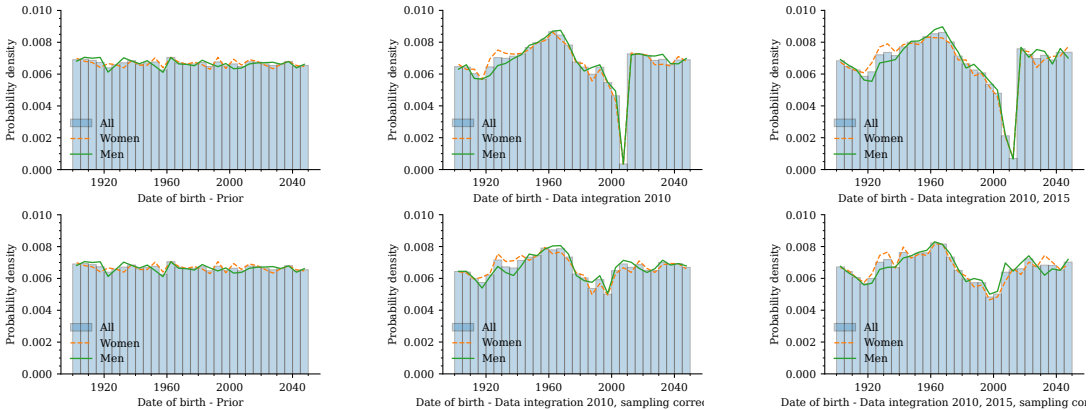
Priors: Projection in 2010





Age and canton of residence in 2010

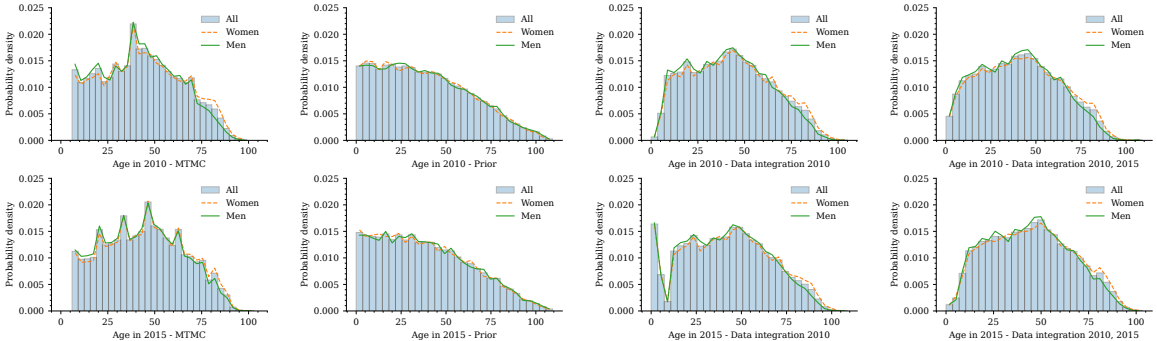
Existence posterior - time independent variables



Date of birth distributions: prior and posterior updates, differentiated by gender

Without sampling correction (first line) and with sampling correction (second line)

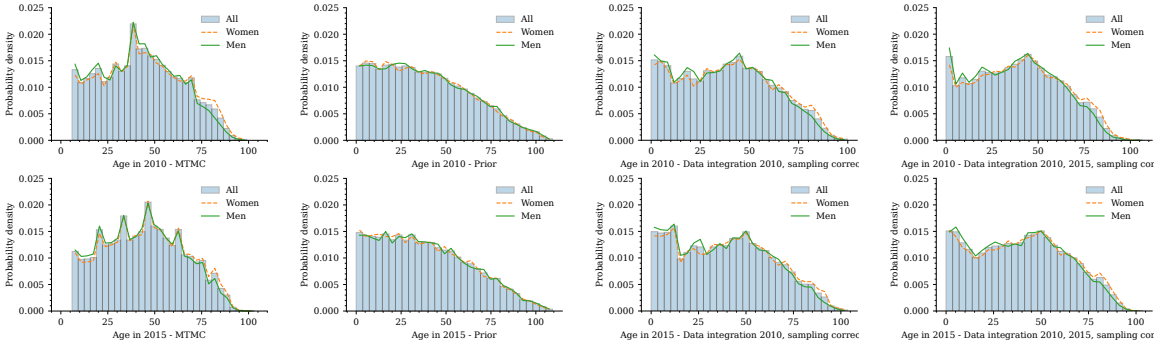
Existence posterior projected - not corrected



Observed data, prior, posterior with 2010, posterior with 2010 and 2015; in 2010 and 2015

No correction of the sampling process of the MTMC

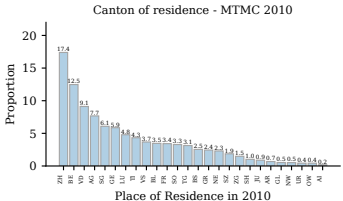
Existence posterior projected - corrected

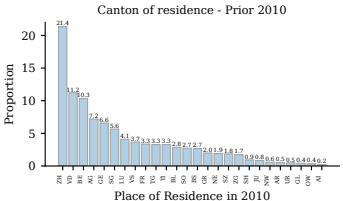


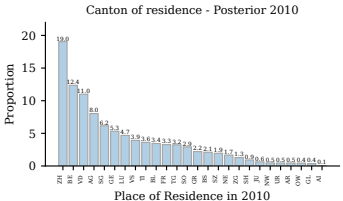
Observed data, prior, posterior with 2010, posterior with 2010 and 2015; in 2010 and 2015

Correction of the sampling process of the MTMC

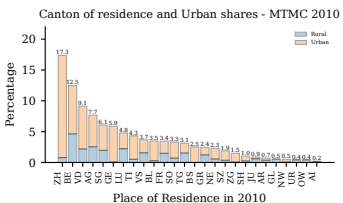
Examples: Place of residence

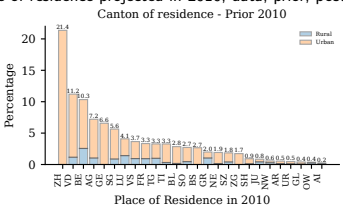


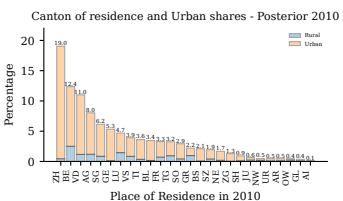




Place of residence projected in 2010; data, prior, posterior







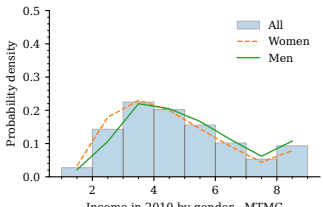
Place of residence projected in 2010 with urban shares; data, prior, posterior

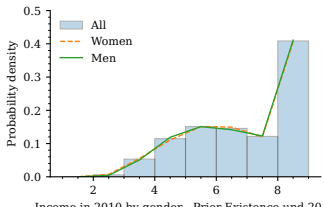
Examples: Employment status

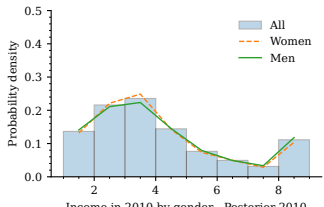
Employment status	**Prior**	**Posterior**	**Data**
Unemployed / inactive	39.19	27.85	13.10
Employed	32.56	43.42	59.11
Retired	18.17	17.95	17.85
Under 15	10.08	10.08	9.92

Table: Employment status proportions in 2010.

Examples: Income

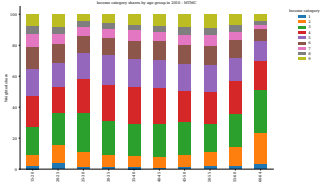


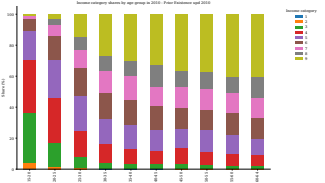


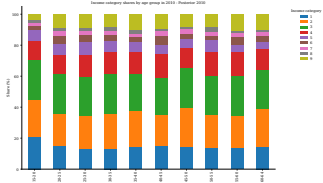


Posterior income distributions in 2010: differentiated by age and gender

Examples: Income







Outline

Conclusion

Current research

- Flexible methodology based on time independent variables.
- Bayesian approach allows to combine models and data.
- Cross-sectional data can be integrated.
- Proof of concept and validation.

Future research

- Synthetic populations of households.
- Integration with activity-scheduling models.

Bibliography I

Abraham, J. E., Stefan, K. J., and Hunt, J. D. (2012).
Population synthesis using combinatorial optimization at multiple levels.
In 91st Annu. Meet. Transp. Res. Board, Washington, DC, USA.

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.

Bodenmann, B. R. and Axhausen, K. W. (2015).
Modeling life-cycle of firms and its effect on relocation choice.
In Bierlaire, M., de Palma, A., Hurtubia, R., and Waddell, P., editors, Integrated Transport and Land Use Modeling for Sustainable Cities, pages 201–218, Lausanne, Switzerland. EPFL Press.

Bibliography II

de Palma, A., de Lapparent, M., and Picard, N. (2015).
Modeling real estate investment decisions in households.
In Bierlaire, M., de Palma, A., Hurtubia, R., and Waddell, P., editors, Integrated Transport and Land Use Modeling for Sustainable Cities, pages 137–160, Lausanne, Switzerland. EPFL Press.

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

Gilbert, C. C. S. (1992).
A duration model of automobile ownership.
Transportation Research Part B: Methodological, 26(2):97–114.

Bibliography III

Gompertz, B. (1833).
On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies.
Proc. R. Soc. Lond., 2:252–253.

Kaldasch, J. (2012).
Evolutionary model of the personal income distribution.
Physica A: Statistical Mechanics and its Applications, 391(22):5628–5642.

Kolvereid, L. (1996).
Prediction of employment status choice intentions.
Entrepreneurship Theory and Practice, 21(1):47–58.

Bibliography IV


Kukic, M., Benchelabi, S., and Bierlaire, M. (2023).
Hybrid simulator for capturing dynamics of synthetic populations.
In 2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC), pages 2646–2651.


Lomax, N., Smith, A., Archer, L., Ford, A., and Virgo, J. (2022).
An open-source model for projecting small area demographic and land-use change.
Geographical Analysis, 54.

Bibliography V


Manzo, G. (2013).
Educational Choices and Social Interactions: A Formal Model and a Computational Test, volume 30 of Comparative Social Research, pages 47–100.
Emerald Group Publishing Limited.


Namazi-Rad, M.-R., Mokhtarian, P., and Perez, P. (2014).
Generating a dynamic synthetic population – using an age-structured two-sex model for household dynamics.
PLOS ONE, 9(4):1–16.

Bibliography VI

Nurul Habib, K. (2018).
Modelling the choice and timing of acquiring a driver's license: Revelations from a hazard model applied to the university students in toronto.
Transportation Research Part A: Policy and Practice, 118:374–386.

Prédhumeau, M. and Manley, E. (2023).
A synthetic population for agent based modelling in canada.
Scientific Data, 10.

Xu, L. and Veeramachaneni, K. (2018).
Synthesizing tabular data using generative adversarial networks.
arXiv:1811.11264 [cs, stat].