Université de Liège

ArGEnCo

CTML
Center for Transport, Mobility and Logistics

LEMA

lepur

Iterative Proportional Fitting (IPF) to expand a synthetic sample into a synthetic population

Mario Cools

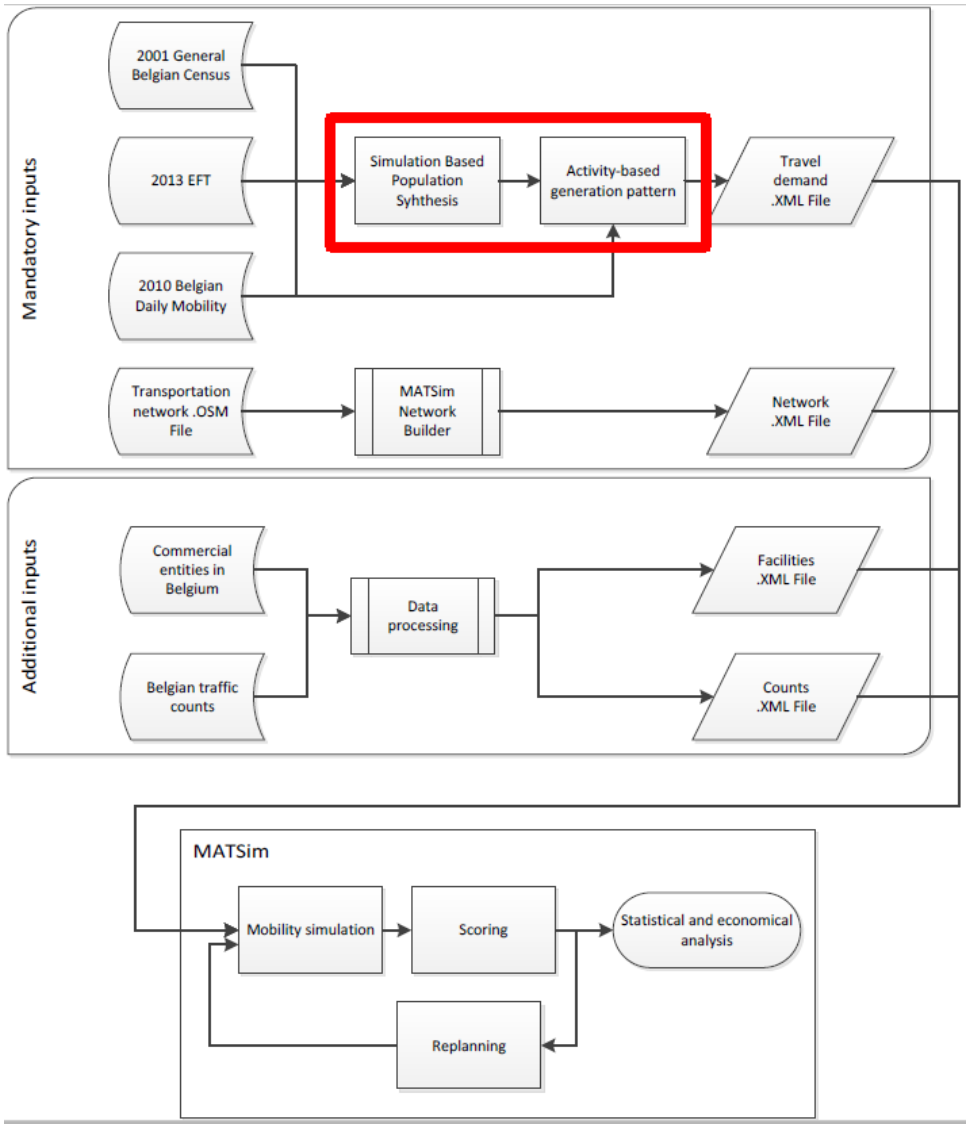
Université de Liège

Outline

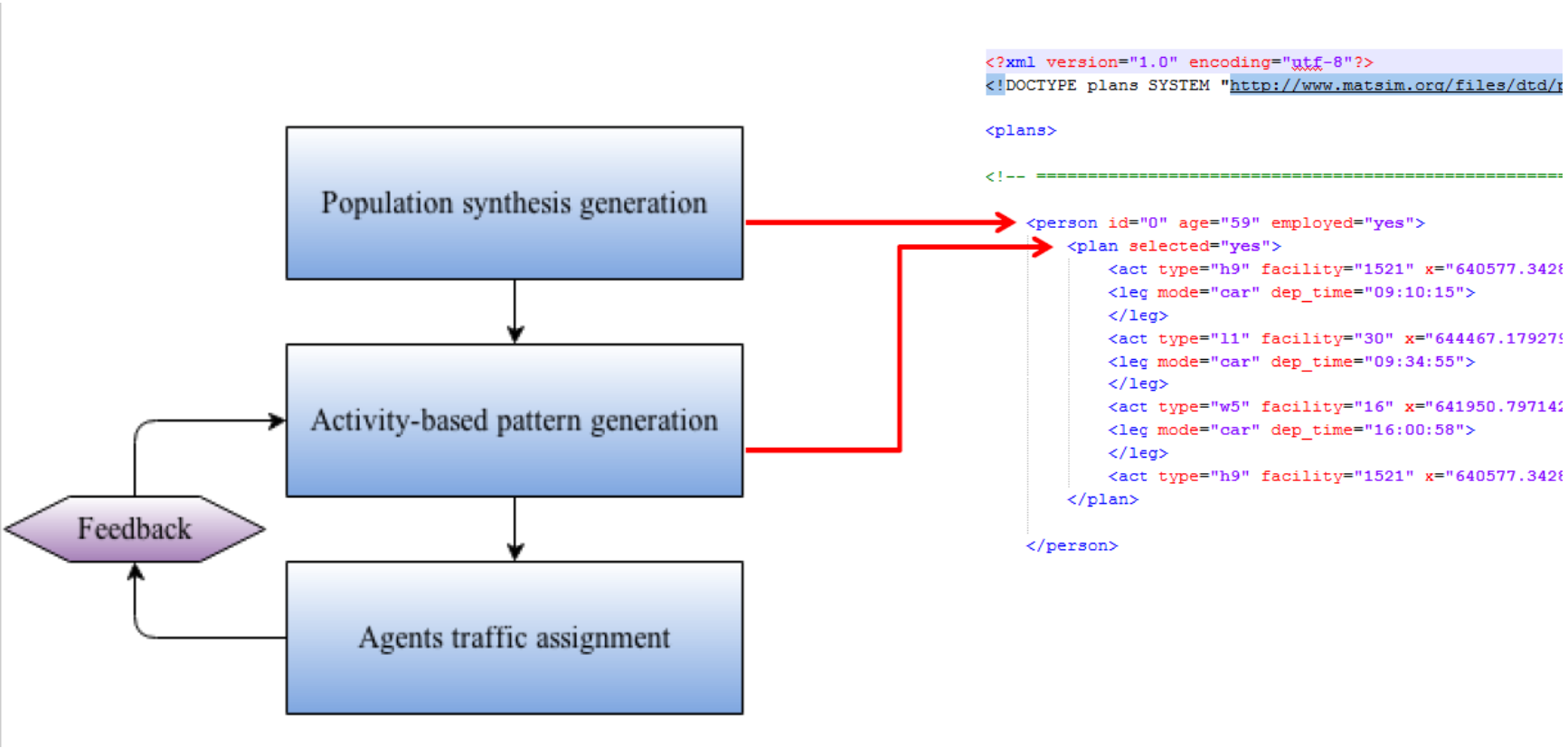
CTML
Center for Transport, Mobility and Logistics

- 1. Context
- 2. Overview of techniques
- 3. IPF (Iterative Proportional Fitting)
- 4. Other results
  - Simulation based approach (MCMC)
- 5. Conclusions
- 6. References

1. Context (1/2)


Mandatory inputs
2001 General Belgian Census
2013 EFT
2010 Belgian Daily Mobility
Simulation Based Population Syhthesis
Activity-based generation pattern
Travel demand .XML File
Transportation network .OSM File
MATSim Network Builder
Network .XML File
Additional inputs
Commercial entities in Belgium
Belgian traffic counts
Data processing
Facilities .XML File
Counts .XML File
MATSim
Mobility simulation
Scoring
Statistical and economical analysis
Replanning

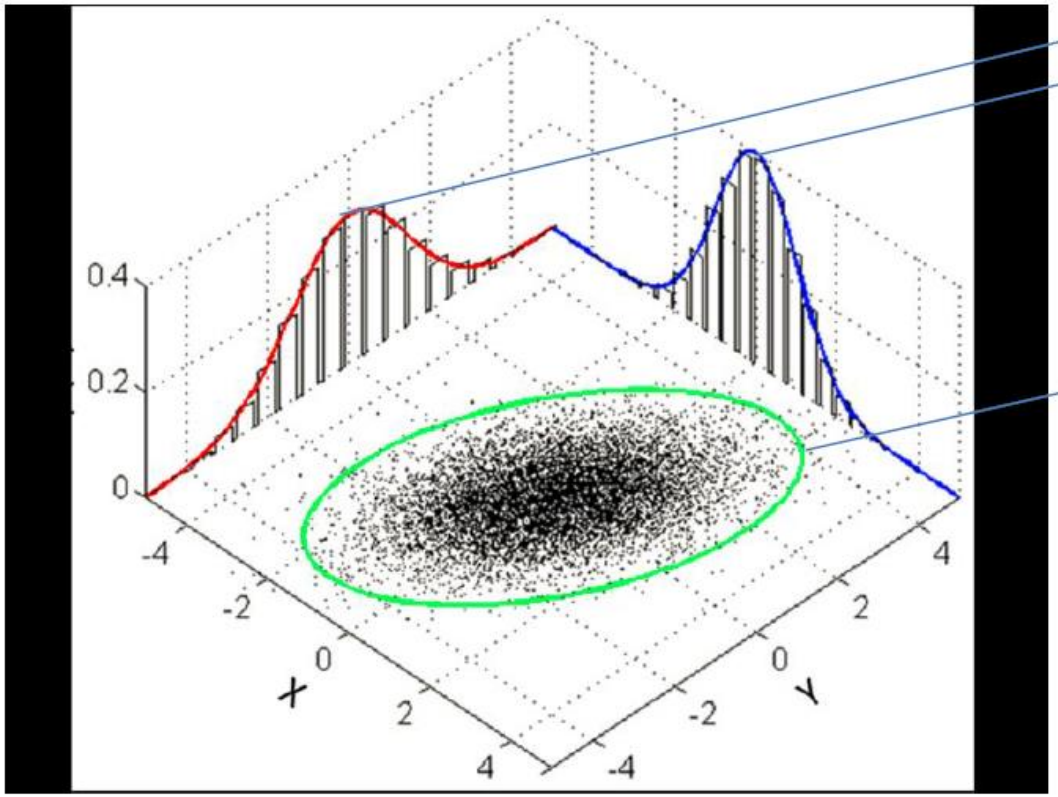
1. Context (2/2)


Population synthesis generation
Activity-based pattern generation
Agents traffic assignment
Feedback
<?xml version="1.0" encoding="utf-8"?>
<!DOCTYPE plans SYSTEM "http://www.matsim.org/files/dtd/
<plans>
<person id="0" age="59" employed="yes">
<plan selected="yes">
<act type="h9" facility="1521" x="640577.3428
<leg mode="car" dep_time="09:10:15">
</leg>
<act type="l1" facility="30" x="644467.179279
<leg mode="car" dep_time="09:34:55">
</leg>
<act type="w5" facility="16" x="641950.797142
<leg mode="car" dep_time="16:00:58">
</leg>
<act type="h9" facility="1521" x="640577.3428
</plan>
</person>

2. Overview of techniques

- Fitting methods
  - E.g. data matching, data fusion, IPF
- Reweighting methods
  - E.g. deterministic reweighting, combinatorial optimization
- Simulation-based methods
  - MCMC (Farooq et al., 2013)

3. IPF (1/8)



Marginal distributions:
Available information

Joint distribution:
initialized by a micro-sample

3. IPF (2/8)

	Y_1	Y_2	...	...	Y_n	*Marginal Probabilities of X*
X_1	p_{11}	p_{12}	...	...	p_{1n}	P_1
X_2	p_{21}	p_{22}	...	...	p_{2n}	P_2
.	.	.	...	...	.	.
.	.	.	...	...	.	.
X_m	p_{m1}	p_{m2}	...	...	p_{mn}	P_m
Marginal Probabilities of Y	P_1'	P_2'	...	...	P_n'	1

Fitting algorithm is applied until meeting the convergence

Problems in terms of convergence
Zero-cell problem

3. IPF (3/8)

Problems in terms of convergence
Zero-cell problem

0,41%	4,07%	8,13%	16,26%	20,39%
0,00%	4,07%	0,00%	0,00%	23,45%
4,07%	8,13%	0,41%	8,13%	20,39%
8,13%	16,26%	20,33%	1,63%	35,78%
13,25%	20,39%	25,48%	40,88%	

3. IPF (4/8)

1. Estimation of a multivariate demographic table of proportions using IPF

2. Drawing of a synthetic population of households from the PUMS in such a way that it matches the proportions of the above estimated table

3. IPF (5/8)

- Four attributes were synthesized,
  - Housing location (20 categories/zones)
  - Age (14 categories)
  - socio-professional status (14 categories)
  - Gender
- Data: BELDAM
  - Assumption: BELDAM true population
  - Samples from BELDAM as micro-sample

3. IPF (6/8)

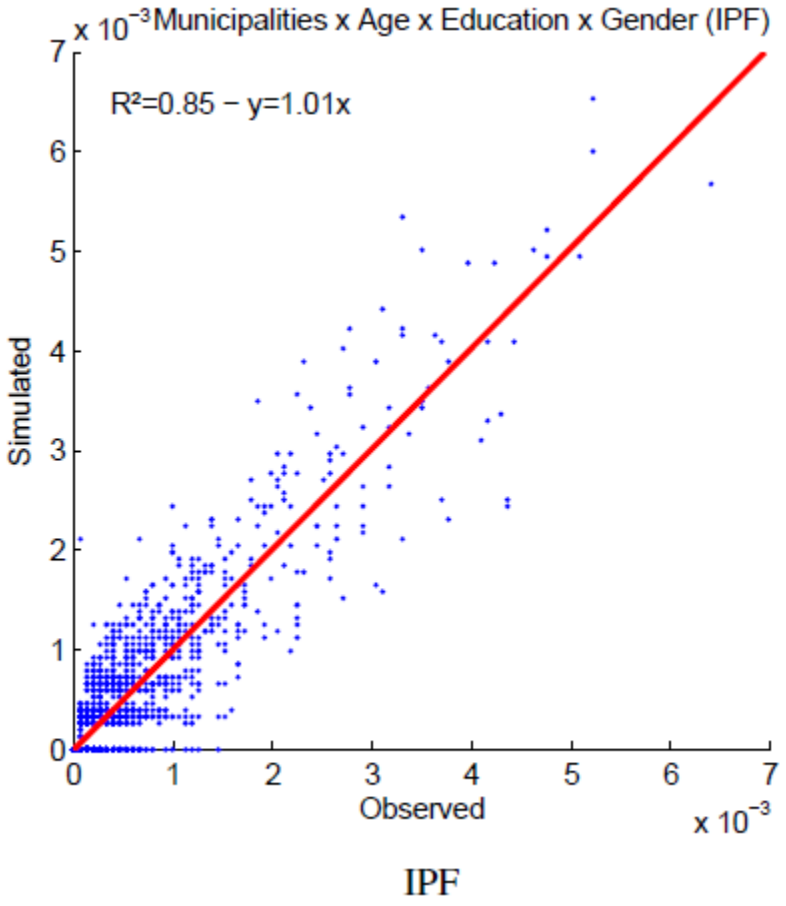
- Goodness of fit:
  standardized root mean square error (SRMSE)

$$SRMSE = \frac{\sqrt{\frac{1}{N}\sum\sum\sum\ldots(\tilde{J}_{ijk\ldots} - J_{ijk\ldots})^2}}{\frac{1}{N}\sum\sum\sum\ldots J^2_{ijk..}}$$

where $\tilde{J}_{ijk\ldots}$ and $J_{ijk\ldots}$ represent the number of agents characterized by the combination of attributes i,j,k,... of the synthesized and observed population, and N is the total number of cells within the matrix.

3. IPF (7/8)

Sampling rate(%)	IPF
1	0.558
2	0.370
3	0.329
4	0.263
5	0.199
10	0.137
15	0.104
20	0.091
25	0.076



3. IPF (8/8)

- Software packages
  - For the analysis: Barthelemy et al. (2015)
    https://cran.r-project.org/web/packages/mipfp/index.html
  - Other:
    http://www.demog.berkeley.edu/~eddieh/datafitting.html

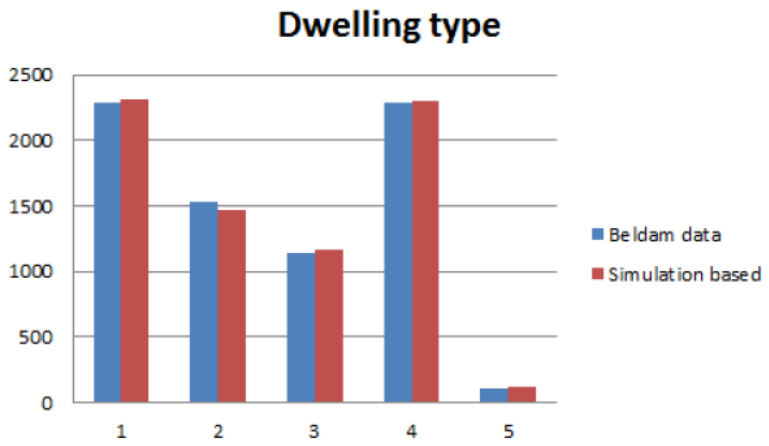
4. MCMC (1/2)

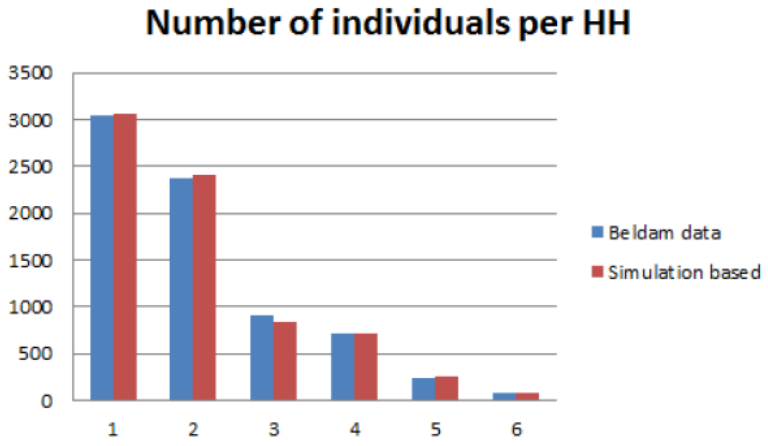
- Approach: Farooq et al. (2013)
- Data used: BELDAM
- 4 covariates
  - Dwelling type
  - Number of cars per HH
  - Number of persons per HH
  - Income level/month per HH
- Full conditionals used
- No dependence with initial condition

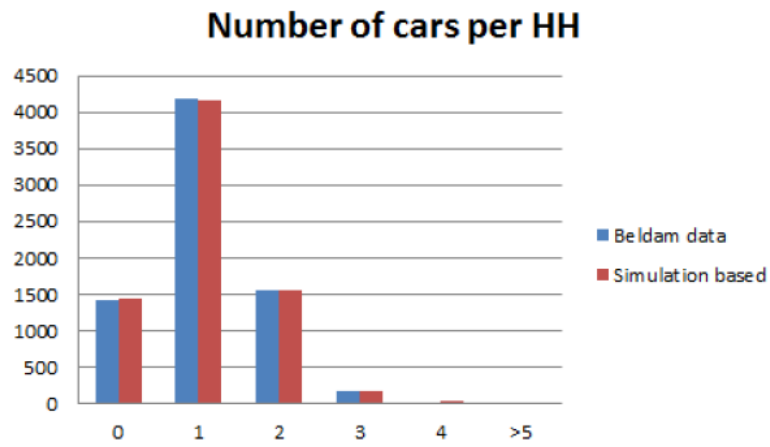
Université de Liège

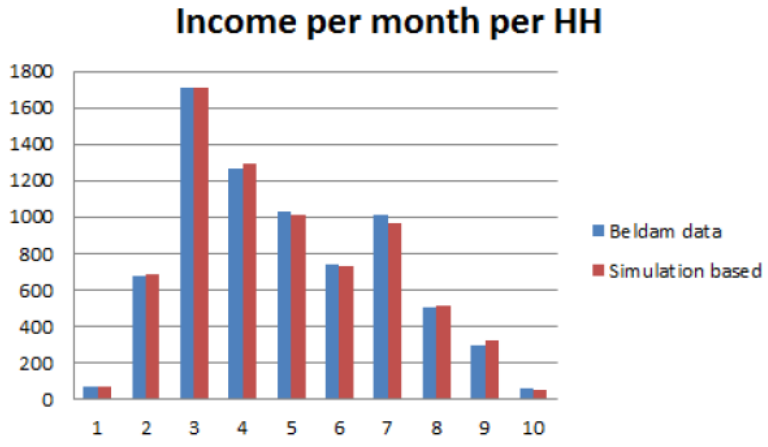
4. MCMC (2/2)

CTML
Center for Transport, Mobility and Logistics









5. Conclusions

- IPF
  - ☒ Can only be used with discrete variables containing a limited number of categories
  - ☒ Fitting a high number of attributes makes the computational process relatively costly
  - ☑ Conceptually easy
  - ☑ Works well for large micro-samples

6. Selected References

- Farooq, B., Bierlaire, M., Hurtubia, R., & Flötteröd, G. (2013). Simulation based population synthesis. *Transportation Research Part B: Methodological* 58, 243–263.
- Saadi, I., Eftekhar, H., El Saeid Mustafa, A. M., Teller, J., & Cools, M. (2014). An agent-based micro-simulation framework to assess the impact of river floods on transportation systems: implementation trajectory for an assessment in the Brussels metropolitan area. *Proceedings of the International Conference on Traffic and Transport Engineering*.
- Saadi, I., El Saeid Mustafa, A. M., Teller, J., & Cools, M. (2016). An integrated framework for forecasting travel behavior using Markov Chain Monte-Carlo simulation and profile Hidden Markov Models. *Proceedings of the 95th Annual Meeting of the Transportation Research Board*. Washington, DC: Transportation Research Board of the National Academies.
- Saadi, I., Liu, F., El Saeid Mustafa, A. M., Teller, J., & Cools, M. (in press). A framework to identify housing location patterns using profile Hidden Markov Models. *Advanced Science Letters*.

Acknowledgements

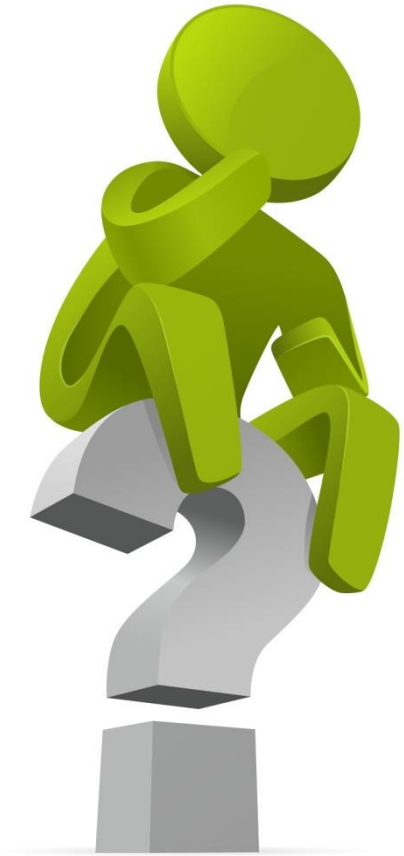
- Ismaïl Saadi

- ARC Project “Floodland” funded by the Fédération Wallonie-Bruxelles
- FSR Project “VEEAT” funded by the Fédération Wallonie-Bruxelles

Université de Liège

Questions?

CTML
Center for Transport, Mobility and Logistics



www.lema.ulg.ac.be
www.ctml.ulg.ac.be