Mobility modeling through mobile data: generating an optimized and open dataset respecting privacy

H. H. Arcolezi, J-F. Couchot, O. Baala, J-M. Contet, B. Al Bouna, X. Xiao

June 18, 2020

Table of contents

1. Introduction

Human mobility modelling:

- Mobile data:
  - ✓ Scientific progress: Urban planning, business development, optimization [1, 2].
  - ✗ Privacy concerns: tracking individuals location, privacy threats (sensitive data).
- Orange Flux Vision® (FV®) [3] publishes statistical indicators respecting privacy wrt european GDPR (General Data Protection Regulation) [4].
- **Objectives:**
  - Privacy of the raw data is critical [5].
  - Improve the utility of these data solving challenges due to data anonymization.
  - Generate a synthetic dataset of Virtual Humans (VHs) to reconstruct the improved mobility scenario.

2. Case study and data analysis

Festival International de Musique Universitaire (FIMU) [6]:

- Modeling n=7 days: Identify people’s mobility patterns in the week of the festival.
- FV® mobility model: Volume of users (resident, french and foreign tourists) per day and per cumulative days (union).

Number of unique users per socio-categorical profile (geoLife)

Date	geoLife	Visitor Category	Cumulative Days	Volume
2017-06-01	popular	French Tourist	6 days	4,000
2017-06-02	NR	Foreign Tourist	2 days	971
2017-06-03	rural worker	Resident	3 days	1,359

2. Case study and data analysis

Challenges

- Data acquisition (many files - volume of users per sensitive data):
  - Generalization (age ranges, socio-professional categories, region, …).
  - *k*-anonymity (indistinguishability among *k*=20 users) -> data masking (#).
  - Extrapolation of Orange's customers to estimate the real population.

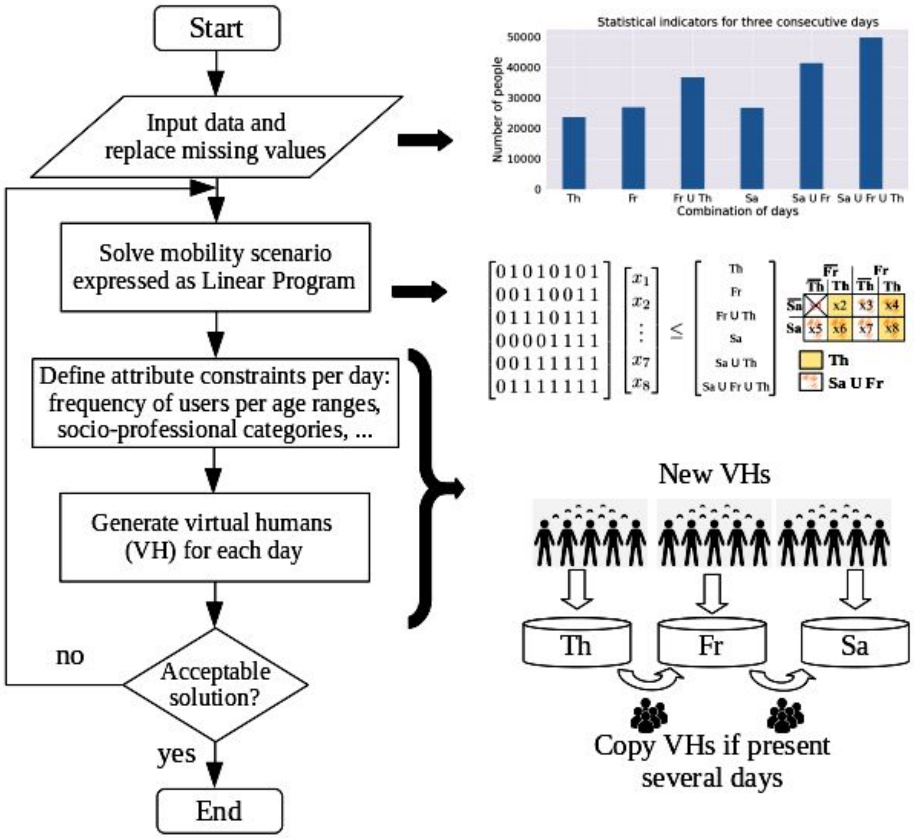
Aggregation of French unique visitors present over three FIMU's days

Label	Cum. Days	FR_geoLife	FR_gender	FR_age	FR_region
Th1	01 day	23,816	23,811	23,810	23,598
Fr1	01 day	27,145	27,144	27,142	26,945
Fr2	02 days	36,917	36,915	36,915	36,758

3. Proposed approach

Random search optimization:

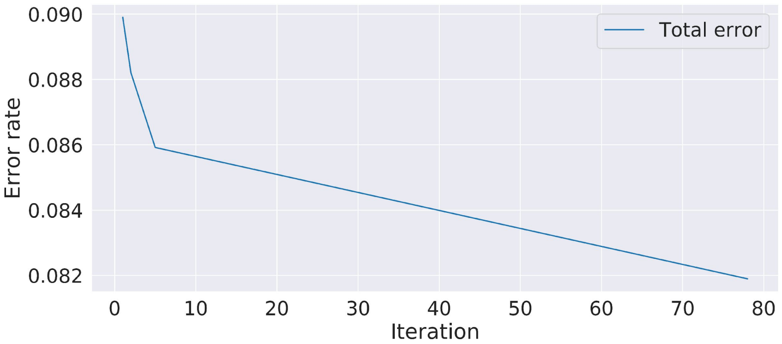
- Input mobility data and randomly replace # with range(1,20).
- Model the improved mobility scenario as a linear program.
- Define user's attributes constraints.
- Generate Virtual Humans (VHs) and define the combination of days constraints.



4. Results: mobility scenario

Improved mobility scenario:

- Allows learning more insights about users' mobility patterns.
- Fast convergence: 100 iterations in 22 minutes (best result found in iter. 79).
- Final dataset: low error rate (~ 8.1%).



Days combination				$\overline{\text{Fr}}$				Fr			
				$\overline{\text{Th}}$		Th		$\overline{\text{Th}}$		Th	
				$\overline{\text{We}}$	We	$\overline{\text{We}}$	We	$\overline{\text{We}}$	We	$\overline{\text{We}}$	We
$\overline{\text{Tu}}$	$\overline{\text{Mo}}$	$\overline{\text{Su}}$	$\overline{\text{Sa}}$	-	4851	4378	1527	1801	1701	786	3450
			Sa	4791	234	87	266	1748	48	417	893
		Su	$\overline{\text{Sa}}$	9695	228	199	508	341	92	506	1220
			Sa	2171	287	74	73	4237	103	1109	1229
	Mo	$\overline{\text{Su}}$	$\overline{\text{Sa}}$	5937	183	49	207	97	36	67	233
			Sa	592	100	103	42	63	116	63	80
		Su	$\overline{\text{Sa}}$	7380	71	34	56	71	89	77	22
			Sa	256	51	96	49	27	52	94	61
Tu	$\overline{\text{Mo}}$	$\overline{\text{Su}}$	$\overline{\text{Sa}}$	7052	446	213	787	1163	35	679	775
			Sa	441	59	104	71	62	94	106	99
		Su	$\overline{\text{Sa}}$	1004	110	53	70	85	87	52	53
			Sa	42	94	50	91	93	38	51	36
	Mo	$\overline{\text{Su}}$	$\overline{\text{Sa}}$	159	309	72	325	442	67	396	94
			Sa	111	76	89	35	71	34	102	434
		Su	$\overline{\text{Sa}}$	434	84	71	41	112	67	89	149
			Sa	4176	61	71	93	211	74	506	**176**

4. Results: synthetic dataset

Final Generated Dataset

Index	Person ID	Date ID	Visit Duration
1	55385	2	6h
2	234	5	4h

Number of users per dataset and absolute error for sub-categories of Age on the 1st FIMU day

Age range	Real data	Synthetic data	Absolute error
18-24	2,312	2,319	7 (0.3%)
35-44	3,230	3,215	15 (0.46%)
>65	3,483	3,439	44 (1.26%)

Sensitive personal data

Person ID	Name	Gender	Age	GeoLife	Vis. Cat.	Region	Sleeping Area
7645	A. Berry	NR	NR	NR	Foreign T.	UK	city of Belfort
10589	A. Maillet	M	<18	rural worker	French T.	Rhône-Alpes	rest of Doubs

5. Conclusion

- Efficiently generates a synthetic dataset with low error rate.
- Few iterations are required to satisfy all user constraints.
- Overcomes challenges due to data anonymization techniques.
- **Future work:**
  - Improve the dataset with more (virtual) sensitive information.
  - Design privacy-preserving algorithms to collect mobile data.
  - Use this dataset to evaluate the designed algorithms.

6. References

[1] T. Kashiyama, Y. Pang, and Y. Sekimoto. *Open PFLOW: Creation and evaluation of an open dataset for typical people mass movement in urban areas*. DOI: 10.1016/j.trc.2017.09.016

[2] V. Caiati, L. Bedogni, L. Bononi, F. Ferrero, M. Fiore, and A. Vesco. *Estimating urban mobility with open data: A case study in bologna.* DOI: 10.1109/isc2.2016.7580765

[3] *Flux Vision: real time statistics on mobility patterns*. orange-business.com/en/products/flux-vision

[4] *General Data Protection Regulation (GDPR)*. gdpr-info.eu/

[5] Y-A. de Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel. *Unique in the Crowd: The privacy bounds of human mobility*. DOI: 10.1038/srep01376.

[6] *FIMU Belfort 2017 : le festival parfait pour bouger à la Pentecôte*. leparisien.fr/culture-loisirs/fimu-belfort-2017-le-festival-parfait-pour-bouger-a-la-pentecote-23-05-2017-6976476.php

Thank you!

The dataset is freely available at -> github.com/hharcolezi/OpenMSFIMU

Contact: heber.hwang_arcolezi@univ-fcomte.fr