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

Héber Hwang Arcolezi, Jean-François Couchot, Oumaya Baala, Jean-Michel Contet, Bechara Al Bouna, Xiaokui Xiao

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

1st Héber H. Arcolezi
Femto-ST Institute,
Univ. Bourgogne Franche-Comté, CNRS
Belfort, France
heber.hwang_arcolezi@univ-fcomte.fr

2nd Jean-François Couchot
Femto-ST Institute,
Univ. Bourgogne Franche-Comté, CNRS
Belfort, France
jean-francois.couchot@univ-fcomte.fr

3rd Oumaya Baala
Femto-ST Institute,
Univ. Bourgogne Franche-Comté, CNRS
Belfort, France
oumaya.baala@utbm.fr

4th Jean-Michel Contet
Orange Business Services
Orange Labs.
Belfort, France
jeanmichel.contet@orange.com

5th Bechara Al Bouna
TICKET Lab.
Antonine University
Hadath-Baabda, Lebanon
bechara.albouna@ua.edu.lbz

6th Xiaokui Xiao
School of Computing
National University of Singapore
Singapore
xkxiao@nus.edu.sg

***Abstract*—Modeling and understanding people's mobility at a temporal and geographical space are very strict requirements for developing better strategies of urban public and private transportation systems as well as establishing improved business techniques. This work proposes a random-search based approach to instantiate statistical indicators through an improved mobility scenario which provides specific information about people attending one or several days for some events. Then, we recreate that scenario with virtual humans, proposing a synthetic and open dataset that matches the original statistical data. The results show the proposed approach is very efficient to model people's mobility, and the generated data has a low error rate compared to the original one.**

***Index Terms*—Mobility scenario modeling, Synthetic data generation, Open data, Inferring anonymous data.**

I. Introduction

There are many interests in studying human mobility patterns in spatial and temporal contexts. For instance, modeling user's mobility is critical for urban planning and business development such as public and private transportation operators, tourism companies, and shopping centres that aim to develop better business strategies. Additionally, if we have prior knowledge about people that stay for few or several days in a certain tourism area, the local companies and the public administration could plan better strategies for guaranteeing better quality services for this population.

Several studies on human mobility show that humans follow particular patterns with a high probability of predictability. Hence, there is a high interest in understanding how people move. However, taking into account users' privacy, research emerges using synthetic and open data to solve such a problem. In [1], the authors provided a approach for creating an open people mass movement dataset. In [2], the authors studied the use of open data for building and validating a realistic urban mobility model. The authors in [3] developed a framework for the generation of individual human mobility trajectories with realistic spatio-temporal patterns. Finally, the authors in [4] proposed a mobility dataset generation method of social vehicles traveling.

In this perspective, Orange Applications for Business (OAB) [5] proposes an innovative solution namely Flux Vision to deliver real-time statistics on attendance and mobility patterns across a territory, a city or a specific area. More precisely, within the territory of interest, OAB solution studies end users connections to mobile network operators, and extrapolates information from retrieved Orange customers to produce statistical estimations of the attending people. Algorithms for data acquisition are compliant with European laws to guarantee the anonymity of each person. Due to the privacy constraints required by the General Data Protection Regulation (GDPR) [6], the acquired data are anonymized before the marketing of information related to the number of people. All identifying attributes are erased or merged to avoid any privacy breach.

Hence, the final form of the data is very suitable for business purposes. One can analyze how often different geographical areas are visited and how many people are moving around this area. The amount of information is for unique visitors per one or more days, which refers to the fact that the same people can be present in only one or more days. However, pursuing scientific studies with anonymized real-life data presents challenges due to the privacy-utility trade-off.

The main objective of this paper is to propose a approach to instantiate a mobility scenario that matches a "sanitized" dataset of mobility. Intrinsically, this dataset is subject to noise resulting from the extrapolation of Orange data and from the anonymization procedure to respect GDPR. First, the proposed approach aims to improve the utility of this data providing more specific information about the attending people with regard to their mobility behavior during days, that is, if people stay a day or more in the same place. Second, it recreates the

This work was supported by the Region of Bourgogne Franche-Comté CADRAN Project.

scenario with virtual humans, such that the synthetic dataset matches the original statistical data. Therefore, as an open dataset, one can carry out studies such as testing and improving data anonymization techniques.

The mobility scenario we propose represents an invaluable source of information to the city public administration and private companies. Rather than being limited to the number of unique people present in certain regions per day, the scenario allows to know if they are the same visitors or different visitors over the analysed time period. With such specific information, companies and public administration would be able to manage their employees and equipment resources efficiently to improve accommodation and transportation systems according to peoples' mobility, thus providing better attendees comfort and security.

The rest of the paper is organized as follow: Section II presents the study case and the data analysis. Section III introduces the proposed approach. Section IV presents the results and its discussion. Finally, Section V provides concluding remarks.

II. Study Case and Data Analysis

In this section, we present the scenario in which OAB collected the data. Additionally, we briefly describe the structure of the original data. We also highlight some challenges one can face working with real-life anonymized data.

A. Study Case

The approach applies to an OAB published database, based on real data with specific information on attending people on the geographical area of the international music festival a.k.a "Festival International de Musique Universitaire" (FIMU). The FIMU is organized and financed by the City of Belfort, France, with the support of student associations, its 31^{st} edition occurred on the first five days of June 2017 [7].

Modeling people's mobility in such events is of great importance for public administration and private companies. Hence, we propose to model a more precise mobility scenario including one day before the FIMU event, the five days of the FIMU, and one day after the FIMU end. In other words, this 7-days scenario for a 5-days event provides information for these institutions to know the number of people who got in and out of the zone of analysis before, during, and after the event.

B. Data analysis

The database at our disposal has seven different files. Among them, five files describe for each day, the number of unique visitors on the last n days, where n ranges from 1 to 7 days. These files are labeled from now on as FO_country, FR_geo, FR_gender, FR_region, and FR_age, where 'FO' stands for foreigners and 'FR' stands for French citizens.

The term 'visitors' is used to define people present at least 1 hour between 06:00 and 23:59 of a given day of the reporting period in the area of interest. In each file relating to French citizens, people are grouped according to their visitor category. "Resident" are people whose billing address is the administrative area around the FIMU. "French tourist" are people billed in France but not in the aforementioned category. The FO_country file has only people grouped as "Foreign tourist" who are people with a foreign mobile phone operator.

TABLE I
Number of unique visitors per geoLife present on FIMU's day.

Date	geoLife	Visitor category	Cumulative days	Volume
2017-06-01	popular	French tourist	6 days	4,000
2017-06-03	NR	Foreign tourist	2 days	971
2017-06-05	rural worker	Resident	3 days	1,359

Moreover, each file classifies people according to the cumulative count from 1 to 7 days, and also by specific categories, which are briefly detailed below[1]:

1) The FR_gender file contains 3,776 rows at total and distinguishes the people by gender (male, female, Not Registered or 'NR'). Furthermore, during the analysis, we noticed very few differences in the frequency of men and women per day (about 50% for both). Hence, in this study, 'NR' values were half assigned for each of both categories, with the same distribution;
2) The FR_age file contains 8,820 rows at total and groups the visitors by age groups;
3) The FR_geo file contains 14,989 rows and groups the visitors in a specific category namely geoLife, divided into different socio-economic sub-categories (e.g., 'rural worker', 'popular');
4) The FR_region file contains 50,350 rows and groups the visitors in the specific category namely 'Region'; there are 22 regions in France;
5) The FO_country file contains 10,832 rows and groups the foreign visitors by country.

For instance, Table I exhibits 3 random samples to illustrate how the volume data are grouped by geoLife profiles in the FR_geo file. Furthermore, there are two additional files labeled from now on as Nights_actual and Presence_time. Unlike previous data files, these latter files do not consider cumulative days information, but the volume of visitors each day $(n = 1)$. Similarly, both files classify the data by the main categories (Resident, French tourist, Foreign tourist) and by specific categories described below:

- The Nights_actual file has 1,145 rows describing for each day the number of visitors who spent a night at the relevant date. Here, people are grouped by a specific category namely 'sleeping area' comprising a dozen areas near the city of Belfort where people spent the night;
- The Presence_time file has 1,301 rows describing for each day the number of hours where visitors were present in the area of interest. Here, people are grouped by a specific category namely 'visit duration' within several sub-categories, for instance, 'Duration 2h' matches people present between one and two hours.

[1]For a complete description of the data, the final results of this research, and to access the synthetic open dataset, the reader can visit the Github page (https://github.com/hharcolezi/OpenMSFIMU).

Note that in the Nights_actual file, the total volume of visitors per day is much less compared to the previous five files (around 4,000 on average). This means that many people did not spend the night near the city of Belfort. Therefore, considering the number of visitors per day from all other files and those in Nights_actual, the term 'NR' was assigned to people that did not sleep in the area of interest.

As stated by OAB, algorithms for data acquisition are compliant with the privacy constraints of GDPR. More precisely, when the number of visitors is less than or equal to 20, data are not published but substituted with symbol #. In the literature this technique is known as *k*-anonymity [8] with $k = 20$ in this case. Additionally, the data are generalized to categories (e.g., age groups). Finally, given the number of identified Orange customers, an extrapolation algorithm is applied to estimate the real population. This latter algorithm is a perturbation-based technique to add noise to the true value.

Such anonymization techniques and extrapolation algorithm to hide personal identifiers from visitors provide a good balance to the privacy-utility trade-off for marketing purposes. Although the identity of individuals are protected, this does not prevent the production of accurate statistics about the attending people in the area of interest.

Even though these data are sufficiently good to be marketed, conducting scientific studies using this data leads to two challenges. First, we are unable to determine the number of people from # values. Instead of excluding this information, these values were randomly replaced by an integer ranging from 1 to 20. Second, the extrapolation algorithm generates an inconsistency between files that describe the same people. Although randomly replacing # values might probably approximate the real value, throughout the analysis of the data that describe the same people, there is a different cardinality between files.

For instance, Table II summarizes the records of the first three days of the FIMU. In this scenario, the first day of analysis is Thursday and has only one record (labeled as 'Th1'), Friday has two records (labeled as 'Fr1' and 'Fr2' respectively), and Saturday has three records (labeled as 'Sa1', 'Sa2', and 'Sa3' respectively). The key points to understand this table are summarized as follows:

1) Both 'Label' and 'Cum. days' columns depend on the date information. The 'Cum. days' column denotes the days that accumulate the number of unique visitors present during one up to three days. For example:
   − The rows which 'Cum. days' value equals 01 day (e.g., 'Th1' or 'Fr1') relate to the information of unique visitors estimated only in their respective day;
   − The rows which 'Cum. days' value equals 02 days (e.g., 'Fr2' or 'Sa2') relate to the information of visitors estimated on their respective day or one day before;
   − The rows which 'Cum. days' value equals 03 days (e.g., 'Sa3') relate to the information of visitors estimated on its respective day, or one or two days before. Similarly, without loss of generality, this analysis can be extended to n days.

TABLE II
FRENCH UNIQUE VISITORS PRESENT OVER THREE FIMU'S DAYS.

Label	Cum. days	FR_geo	...	FR_region
Th1	01 day	23,816	...	23,598
Fr1	01 day	27,145	...	26,945
Fr2	02 days	36,917	...	36,758
Sa1	01 day	26,894	...	26,699
Sa2	02 days	41,615	...	41,373.
Sa3	03 days	50,024	...	49,823

2) The 'FR_geo' and 'FR_region' columns present the total number of unique French visitors aggregated in each file. This is according to the 'Cum. days' attribute and after replacing all # values. Actually, data can be aggregated using 'Date' and 'Cum. days' attributes from Table I. Similarly, without loss of generality, the same procedure is reproduced for the other files.
3) Theoretically, the information from both columns 'FR_geo' and 'FR_region' should be equal as they describe the same population. However, due to the privacy-utility trade-off aforementioned, that is not true. The difference between file changes depending on the replacement of all # values.

III. PROPOSED APPROACH

Our goal is to improve the understanding of people's mobility behavior from the number of unique visitors per day and cumulative days. First, we propose a mobility scenario modeling approach to provide more specific information about people. Second, we provide a scheme to generate a synthetic dataset based on the resulting mobility scenario.

The whole proposed approach is summarized with a flowchart depicted by Fig. 1. In this particular study, the ultimate goal is to infer the number of people who stayed in the city for one or any combination of days considering one week including the FIMU event. Further, once the whole mobility scenario is known, the objective is to generate samples to build a synthetic dataset with virtual people. The approach is detailed and applied in the following two subsections.

A. Mobility scenario modeling

As described on the left side of Fig. 1, we input data with cumulative information and replace the # values. A Boolean map is used to describe every combination of $n = 7$ consecutive days resulting in $2^n = 128$ variables. Then, each of the $n(n + 1)/2 = 28$ cumulative days is described as a Boolean vector with 0 (excluded) and 1 (included) values per combination of days according to the representative map.

Then, a linear program (LP) is defined to instantiate the first feasible solution given a random initial solution, without trying to maximize or minimize any combination of days. The system constraints are the number of people per cumulative days, expressed as Boolean vectors. However, due to both problems of # values and inconsistencies between the cardinalities of the datasets, rather than using the exact 'known values', these problems are addressed by establishing bounds. The motivation for such an approach is to instantiate a feasible solution that respects the values of all available data the global

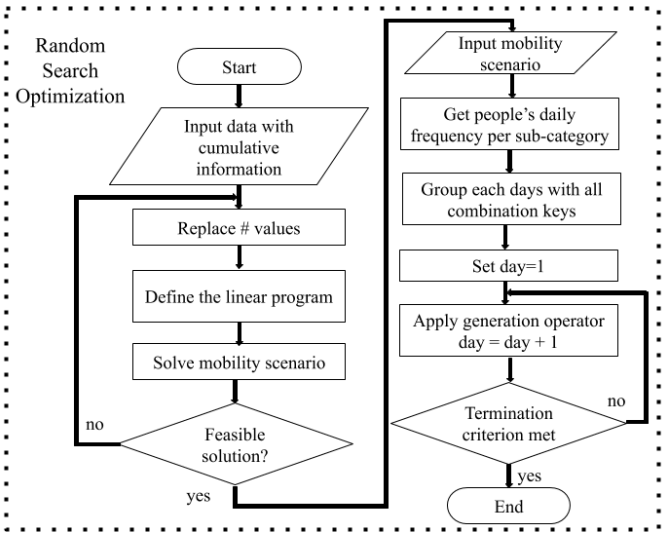

Fig. 1. Flowchart to the proposed algorithm.

error could minimize. More precisely, we are defining a linear constraint solver that computes an arbitrary solution within the set of feasible solutions rather than using the linear program as an optimization mechanism. In this case, the objective function of this system is just a constant (zero).

In this context, there are two scenarios. The first one is for French citizens, where the lower and upper bounds are the minimum and maximum value between all the datasets. The second one is for foreign tourists where the lower and upper bounds are the own value encountered after randomly replacing # adding and removing 20 respectively. Equation (1) mathematically describes the LP as:

$$\begin{aligned} &\min 0, \\ &\text{s.t. } lb_i \leq A_{ij}x_j \leq ub_i \text{ and } x_j \geq 0 \end{aligned} \tag{1}$$

$\forall i \in [1, n(n+1)/2]$ and $\forall j \in [1, 2^n]$ where A_{ij} is the Boolean matrix representing the Boolean vectors i and its respective days combinations j, x_j is the number of people per combination of days. And, lb and ub are both lower and upper bounds respectively.

Hence, instantiating a feasible solution for all the categories (Resident, French tourists, and foreign tourists) and grouping them as a unique mobility scenario provides the number of people for each combination of days. Then, with such results, the second part is retaken for generating samples of virtual humans aiming to approximate the original data.

To better understand the proposed LP, let us consider the scenario of Table II. Fig. 2 illustrates the Boolean map representation of $n = 3$ consecutive days (Th=Thursday, Fr=Friday, Sa=Saturday, and its complements), and the example of both $Th1$ and $Sa2$ variables (unique visitors on Thursday and unique visitors present on Saturday or Friday). Note that the block indicating the variable $x1$ represents all the previous and subsequent days besides those of the analysis, hence it is not considered.

Considering the LP in Equation (1), Equation (2) illustrates the A_{ij} matrix according to Fig. 2 and its upper bound (ub) with values from Table II relating to French citizens.

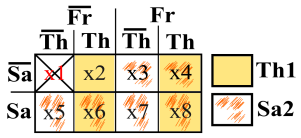

Fig. 2. Representation of n=3 days combination and illustration of both Th1 and Sa2 known values.

$$\begin{bmatrix} Th1 \\ Fr1 \\ Fr2 \\ Sa1 \\ Sa2 \\ Sa3 \end{bmatrix} = \begin{bmatrix} 0\,1\,0\,1\,0\,1\,0\,1 \\ 0\,0\,1\,1\,0\,0\,1\,1 \\ 0\,1\,1\,1\,0\,1\,1\,1 \\ 0\,0\,0\,0\,1\,1\,1\,1 \\ 0\,0\,1\,1\,1\,1\,1\,1 \\ 0\,1\,1\,1\,1\,1\,1\,1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_7 \\ x_8 \end{bmatrix} \leq \begin{bmatrix} 23,816 \\ 27,145 \\ 36,917 \\ 26,894 \\ 41,615 \\ 50,024 \end{bmatrix} \tag{2}$$

B. Synthetic data generation

The proposed algorithm illustrated on the right side of Fig. 1 is summarized in the subsequent steps. First, using the original data, the frequency of visitors present each day of the week under study is calculated for each sub-category, e.g., on the first day 50.2% are men and 49.8% are women.

We set up a dictionary for each day grouping its related keys of combination days; people present only Thursday are described by TT, people present both Thursday and Friday are described by TF, and so on. It is noteworthy that the same TF key appears on both Thursday and Friday dictionaries as they are the same people that attended both days in the analysis area.

An iteration starts filling up each key for the first day with virtual individuals respecting the frequency of men and women, geoLife categories, age groups, regions (countries for foreign tourists), the sleeping area, and the visit duration. Afterward, for the next six days, people with similar keys are directly copied from one day to the another. In this case, the frequency for each category is re-calculated considering the existing people. The remaining people are then generated according to the new frequency. However, there is one exception about the attribute 'visit duration', which means that people could be present more hours from one day to another. Hence, the approach may vary the duration time of every people each day relative to the real frequency acquired from the original data.

Once the stop criterion is met, when all days have their respective virtual humans, the error is calculated by querying the generated data and comparing it to the original one. The error, total error, and error rate metrics are defined in the following.

Definition 1 (Error)**.** *Let $|A|$ be the cardinality of set A and $A_{|k}$ be the subset of A restricted to sub-category k, i.e., $A_{|k} = \{x|x \in A, x \in k\}$. Given a set O (original data), a set G (generated data), and sub-categories k related to each specific category (i.e., from the gender category there are two sub-categories, feminine and masculine), the error is defined as*

$$error(k) = ||G_{|k}| - |O_{|k}||$$

Definition 2 (Total Error)**.** *The total error TE is the sum of errors per sub-category k and per day i defined as*

$$TE = \sum_{i=1}^{n} \sum_{k=category} error(k)_i$$

Definition 3 (Error Rate)**.** *The error rate ER is calculated considering j original datasets*

$$ER = \frac{TE}{\sum_{j=dataset} \sum_{k=category} |O_{j|k|}|}$$

These computations are repeated for m iterations based on a random search approach. In particular, the first parameter randomly generated is the # values within the range 1-20, which changes the people number per day of each file at every iteration. Considering the LP in Equation (1), an initial solution is randomly generated such that the linear constraint solver can provide a different mobility scenario at each iteration. Then, the error rate metric is calculated. Finally, the best mobility scenario and synthetic datasets are recovered as a final solution.

The motivation for such random search approach is as follows. First, an initial attempt to model our problem as a linear program resulted in infinity solutions. Second, as aforementioned, the # problem due to privacy constraints had to be handled resulting in different cardinalities for the datasets. Third, using random search has fewer computational costs than grid search, and the final solution is of good quality.

IV. Results and Discussion

To carry out this work, we used the Pyeda python package [9] for Boolean algebra operations. In order to run the codes, we used two machines that have the following characteristics: one with a "Titan X", Intel(R) Xeon(R) CPU E5-2623 v4 @ 2.60GHz, 64Gb of RAM, and a GPU with 3,072 cores and 12Gb of RAM. The other with a "3 Titan V100", Intel(R) Xeon(R) Silver 4110 CPU @ 2.10GHz with 128Gb of RAM, and each "Titan V100" with 5,120 cores and 16Gb of RAM. We applied the mixed-integer nonlinear programming (MINLP) solver from the Gekko package [10] to the proposed mobility scenario in Equation (1). The Faker package [11] assigned fake French names for French citizens and default names (United States) for foreign tourists. In the next two subsections, we present our results.

A. Mobility scenario

The random search algorithm performs $5,000$ evaluations of $m = 100$ iterations in parallel to search for the most representative distribution of people over the week of interest. This is a suitable way to ensure convergence pattern towards a global minimum due to probabilistic properties. At the end of 22 minutes, the random search stops, and the dataset is recovered with an error rate less than 8.1% at evaluation $1,050$ and iteration 79.

We summarize the best result describing the proposed mobility scenario in a large table that can be hardly integrated into the text due to space constraints. We invite the reader to refer to footnote[1] in Subection II-B. On one hand, Fig. 3 depicts the error rate decreasing function based on the number of iterations. On the other hand, Table III presents the number of visitors for both real and synthetic datasets (FR_age) and the absolute error for three sub-categories of ages on the first day of interest.

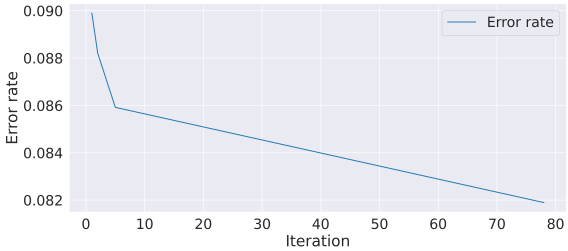

Fig. 3. Change in error rate through iterations.

TABLE III
NUMBER OF VISITORS PER DATASET AND ABSOLUTE ERROR FOR EACH SUB-CATEGORY OF AGES ON THE FIRST DAY.

Age group	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%)

It is noteworthy that the approach described in Section III and summarized in Fig. 1 considers a virtual person having the same categories attributed in the beginning, then copies it to other days if present. This naturally means that each person in that population has always the same geoLife profile: people are from one unique region, they normally sleep in the same zone, and so on, except for the 'visit duration' attribute.

Hence, as noticed in Fig. 3 and Table III, the error metrics are very low when querying people in each sub-category from the generated data, compared to the original one. In other words, the result which is one of many possible scenarios, closely describes how people behave during the week of interest. With such results, it is possible to assert with a reasonable amount of accuracy how many people were present each combination of 7 days, which is a more precise mobility scenario than just knowing the number of unique people per day or cumulative days.

From the final mobility scenario, and by querying the generated dataset, we can find out how many foreign tourists, French tourists and residents are present only one or several days at FIMU event, as well as their specific information such as socio-category profile, region or countries, age groups, gender, and so on. For illustrative purposes, it is possible to know that from 176 visitors present during all week, 153 are residents, 20 are French tourists, and 3 are foreign tourists.

Moreover, it is noticed that foreign tourists were present normally at one unique day or at most three consecutive days, which is consistent with reality. Indeed, foreign people come to the FIMU for few days and have no 'gaps' between days, such as one-day present, the other not, and the next yes. Additionally, the premise of assigning one unique sleeping area for each visitor proves that the approach is consistent.

Such specific information is valuable for local communities,

TABLE IV
FINAL GENERATED DATASET.

Index	Person ID	Date ID	Visit Duration
1	5385	2	6h
2	234	5	4h

and for accommodation and transportation companies, and allows them to learn how people behave during a time period in a particular area. Indeed, if we have information about the presence of foreign tourists on a specific combination of days and if they do not change much their sleeping place, accommodation companies can improve their future strategies to assist this population. Similarly, tourism companies would be more prepared knowing that most of people present during the week are residents. Instead, tourists are rather present during the weekend of the FIMU event.

B. Synthetic data

In the end, an open dataset is proposed with an associative table whose primary key is (Person ID, Date ID) combination, which specifies the visit duration information, as shown in Table IV. These two individual keys are linked to two other tables. The first table contains specific information about people, for instance, fake French names, geoLife profile, and region. The second table maps the days under analysis, from the first to the last day respectively as follows: {1: 2017-05-31}, ..., {7: 2017-06-06}.

The motivation to release the synthetic open dataset is to facilitate its improvement. The ultimate goal is to provide more specific virtual information about people to test different anonymization techniques through synthetic inferred data from the real-world. Therefore, the associative table will remain unaltered, while more attributes can be added to the table with specific information about people. The generated dataset is available for anyone to freely access, use, modify, and share for any purpose at the Github page referred in the footnote[1] in Subection II-B.

V. CONCLUSION

With the increasing of massive data generated by mobile phones, the acquisition of mobile data is of great interest and presents a major issue. Indeed, mobile data availability provides a means to understand people's mobility behavior and patterns. Local authorities, public administrations and private companies can take advantage of such knowledge to identify strategies and make decisions about new transport policies, rearrangement of security in certain regions, to propose better infrastructures and services for the community or attendees to some events.

With this in mind, this paper proposes an approach to infer and recreate synthetic data that provides a precise mobility scenario based on one-week statistical data made available by [5]. Improved mobility scenario presents specific information about people present on one or several combinations of days. This is a way to allow local authorities and private companies re-organize the strategies of existing markets or new ones.

The approach is generic enough to apply to other mobility scenarios that rely on databases with information about the cumulative number of unique people for days. Moreover, the proposed approach overcomes challenges due to data acquisition with anonymization techniques, such as generalization (e.g., age ranges), k-anonymity ($k = 20$ in this case), and extrapolation algorithm.

The results show that the proposal can be efficiently applied to generate a synthetic dataset with specific information about people present in a certain region, for instance, attending the FIMU as was the case in this study. Finally, the generated dataset closely matches the original one with a low error rate, which substantiates the proposed approach. In future work, we aim to continue improving the synthetic dataset by enriching data collection with more specific information about people.

ACKNOWLEDGMENT

Computations have been performed on the supercomputer facilities of "Mésocentre de Calcul de Franche-Comté". The authors would also like to thank the OAB team for their useful feedback and comments.

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," *Transportation Research Part C: Emerging Technologies*, vol. 85, pp. 249–267, Dec. 2017. [Online]. Available: https://doi.org/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," in *2016 IEEE International Smart Cities Conference (ISC2)*. IEEE, Sep. 2016. [Online]. Available: https://doi.org/10.1109/isc2.2016.7580765

[3] L. Pappalardo and F. Simini, "Data-driven generation of spatio-temporal routines in human mobility," *Data Mining and Knowledge Discovery*, vol. 32, no. 3, pp. 787–829, Dec. 2017. [Online]. Available: https://doi.org/10.1007/s10618-017-0548-4

[4] X. Kong, F. Xia, Z. Ning, A. Rahim, Y. Cai, Z. Gao, and J. Ma, "Mobility dataset generation for vehicular social networks based on floating car data," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 5, pp. 3874–3886, May 2018. [Online]. Available: https://doi.org/10.1109/tvt.2017.2788441

[5] Orange-Business-Services, "Flux vision: real time statistics on mobility patterns," https://www.orange-business.com/en/products/flux-vision, 2013, online; accessed 25 September 2019.

[6] European-Commission, "2018 reform of EU data protection rules," https://gdpr-info.eu/, online; accessed 25 September 2019.

[7] A.-C. Dancourt, "FIMU belfort 2017 : le festival parfait pour bouger à la pentecôte," http://www.leparisien.fr/culture-loisirs/fimu-belfort-2017-le-festival-parfait-pour-bouger-a-la-pentecote-23-05-2017-6976476.php, 2017, online; accessed 05 November 2019.

[8] L. Sweeney, "k-anonymity: A model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 05, pp. 557–570, Oct. 2002. [Online]. Available: https://doi.org/10.1142/s0218488502001648

[9] C. Drake, "Pyeda," https://github.com/cjdrake/pyeda, online; accessed 25 September 2019.

[10] L. Beal, D. Hill, R. Martin, and J. Hedengren, "Gekko optimization suite," *Processes*, vol. 6, no. 8, p. 106, 2018.

[11] D. Faraglia, "Faker," https://github.com/joke2k/faker, 2012, online; accessed 25 September 2019.