Using Mobility Profiles for Synthetic Population Generation

Alperen Bektas[1] and René Schumann[1]

SILAB - Smart Infrastructure Laboratory
University of Applied Sciences Western Switzerland
Rue de Technople 3, Sierre, Switzerland
{alperen.bektas | rene.schumann}@hevs.ch

Abstract. Agent-based modeling (ABM) is a wide-spread technique that can be utilized as an artificial laboratory for in-silico experiments of real-case studies of different domains such as mobility. To initialize agent/environment attributes and their relationships, disaggregated (individual level) micro-data is required as an input. However, having such data is not often possible due to several reasons such as privacy concerns. To bridge the gap, generating realistic synthetic data (from census/survey data) becomes an initial and essential step of agent-based modeling. In this piece of research, we employ the mobility profiles of the Swiss population for generating synthetic populations along with their mobility activities. To validate the synthetic data, an agent-based model, which is already calibrated to the empirical data, is re-run with a sample and the generated synthetic data. Accumulated decisions of agents in both cases are compared. In addition, marginal frequencies of control attributes are benchmarked. The first obtained results demonstrate that increasing size of the generated population decreases the difference between simulation results of the synthesized data and the real data.

Keywords: Synthetic population generation · Agent-based modeling · Mobility profiles · Cluster analysis · Demand modeling

1 Introduction

In recent years, the demand for methods/paradigms, which require individual level data (micro-data), has significantly risen. This increase can be attributed to various determinants such as developments in computational power, easier data collection or increasing trend of capturing heterogeneity. One of these paradigms is agent-based modeling (ABM). It is considerably useful to reflect individual behaviors as well as environmental and demographic attributes. It provides a flexible platform where various disciplines (e.g. sociology, psychology, computer science) blend together as one. Agent interactions with each other and with the environment, decision-making mechanism and collective outcomes can be modeled by using theories of these disciplines. Thus, agent-based models can be used as artificial laboratories for in silico experiments of real-case studies.

Through a calibrated and validated agent-based model, future scenarios can be explored before implementation.

To depict real world dynamics in a bottom-up designed agent-based model, disaggregated (individual level) micro-data is required. Except for a few exceptions such data is not often available due to several reasons such as privacy concerns [6]. Even if such data is available for one-to-one matching with agents, this is not favorable because of the strong dependency to the data source. Due to these reasons, generating realistic synthetic populations of simulated areas has become an initial and essential step for agent-based models. A synthetic population is a microscopic representation of a real population [1]. It is not one-to-one identical to the real population. Instead, it mimics, regarding some specific attributes by having similar statistical distributions. It is statistically close enough to the real population to be used in models such as an agent-based simulation.

As a case study of agent-based modeling, we are currently undertaking modeling heterogeneous mobility demand by an agent-based model called BedDeM (Behavior Driven Demand Model) that is explained detailedly in Section 4. Rather than routing (exact coordinates), we investigate potential determinants, which in principle can influence mobility behaviors of individuals in particular modal choices (e.g. car, train, soft mobility modes). To make decisions of agents in the model as realistic as possible, we use a decision making mechanism from psychology called Theory of Interpersonal behavior (TIB) [12]. So far we've been using a sample from a joint table of two data sets (Micro-census (MTMC)[16] 2015 and Swiss Household Energy Demand Survey (SHEDS)[17]) to initialize the agent population in our previous experiments. Characteristics of the data are explained in Section 3. Since dependence to real data sets is not favored, in this research we aim to generate realistic synthetic data for our model. Another reason is that being dependent on data-sets does not allow us to expand the number of agents more than the number of respondents (e.g. to have a high resolution). In the previous research, we clustered the Swiss population based on their mobility related features such as the modal choice to obtain mobility profiles [3]. Narrowed down intra-cluster distributions were obtained along with the medoids of each cluster (i.e. profile). Through an optimization process, intra-cluster cohesion and inter-cluster separation were enhanced that leads to less variation (intra-cluster). The idea of this piece of research is generating synthetic data profile by profile according to shrunk intra-profile distributions (i.e. less variation) and then merging them properly to obtain the final synthetic data (see Section 5). As validation, BedDeM simulation is employed with a sample from real (empirical) data and 3 synthetic data sets (with different population sizes) separately. Simulation results of the real and the synthetic data are compared in Section 6. Besides, marginal frequencies of control attributes are benchmarked. In the next section, we explain some other related studies. The paper ends with limitations and conclusion sections.

2 Related Work

There are various approaches for synthetic population generation. This study is related to most of them in terms of the problem rather than the methodology. Two of the approaches Synthetic Reconstruction (SR) and Combinatorial Optimization (CO) come to the forefront with their variations [9].

SR methods consist of two steps; fitting and generation. These methods are generally based on Iterative Proportional Fitting (IPF) technique (it was first established by Deming and Stephan (1940) [5]), that generates a multivariate table of conditional probabilities, which are derived from cross-tabulations of the desired attributes of the base population. In the fitting stage, cells in the table are fitted to sub-totals that are gained through survey/census data. After that, in the generation step, joint probabilities obtained in the fitting stage are utilized to expand micro level sample data to the full population. Frick's paper gives a good insight into how to use the IPF for categorical variables [7]. In that study, a synthetic population with hectare based level resolution is generated. Farooq et. al. touch on shortcomings of the IPF [6]. The paper states dependency on the sample data is essentially blowing up of the sample rather than reproducing it from the heterogeneous points in the attribute space. Limitation to categorical variables is another shortcoming that the paper taps. It introduces a Markov Chain Monte Carlo (MCMC) simulation-based technique that can overcome the shortcomings. It benchmarks the simulation-based approach with the IPF technique via the standard root mean square error (SRMSE). Even in the worst case, the simulation-based approach overcomes. The technique in the paper is restricted to non-hierarchical data that is underlined by Casati et. al. [4]. They introduce an extended version of MCMC called, hierarchical MCMC (hMCMC) to overcome that restriction. In the paper, hMCMC is combined with generalized ranking (GR). Jeong et. al. also underline the IPF's dependence structure of the reference joint table [11]. The paper introduces a novel capula-based joint fitting (CBJF) approach. It compares the CBJF technique against the IPF. In most cases, the CBJF is superior to the IPF. One of the limitations if the CBJF seems that it can be applied only to ordinal variables. Antoni et. al. use a population synthesizer tool, called MobiSim, which generates agents and distributes them to households according to demographics data [1]. In another study, the performance of two synthesizer tools PopSynWin and PopGen are compared [10]. Both use SR techniques. PopGen performs better in generating population at the individual level. The authors argue that the performance of tools varies according to variation in household and person characteristics of a particular geography. Similar SR based synthesizer tools are addressed such as ILUTE, FSUMTS CEMDAP, ALBATROSS, etc. in other studies [14, 15, 13, 2]. Another comparison research was done by Harland et. al. [8]. The study compares three techniques; deterministic re-weighting, conditional probability, and simulated annealing algorithm. Synthetic populations are generated by them for the city of Leeds in the UK. Simulated annealing was found the best performing one among them.

CO techniques (introduced by Williamson [18]) are concerned with finding an optimal or close to optimal solution among a finite collection of possibilities (joint distribution pools). They involve the random selection of a group of individuals from disaggregated data so that it matches the population size of the small area. Huynh et al. present a paper that can be useful to understand the concept of CO. They use the CO technique to generate a synthetic population of Sydney for their agent-based model [9]. The paper depicts a methodology to initialize and evolve a synthetic population in the model. The initial population is synthesized over aggregated data of demographic distributions and attributes of agents evolve in time endogenously (aging, dying, marriage, divorce, etc.) in the model within a time series (2006-2011).

In brief, there are several validated approaches. We aimed to obtain mobility profiles by clustering in the previous research according to our needs (e.g. exploring mobility profiles, using these profiles to calibrate our model [3]). In this piece of research, we utilize already obtained profiles (i.e. as a milestone) for synthetic population generation. So, we've not used one of the validated approaches that are explained above although some of them are very successful.

3 Characteristics of the Data

In this section, we look close to the real (empirical) data that is utilized by the model so far as to simulate heterogeneous mobility demand in Switzerland. Two qualitative data sources, a census (MTMC [16]) and a survey (SHEDS [17]), were combined to obtain three tables (subsets of the data); the population (agent/individual attributes), the schedule (trips/activities), and the correspondence (vehicle and resource attributions) [12]. This study aims to generate synthetic versions of these tables that mimic the real ones statistically. The population table consists of 180 individual attributes (of respondents) [12]. The table contains mixed-type attributes (i.e. categorical, numeric). The MTMC is utilized to obtain socio-demographic attributes (e.g. age, income level, household size, canton, municipality type, education level, etc.) while the SHEDS is mainly used for psycho-social values (e.g. environment friendliness, mobility preferences, habits, emotions, etc.) to map the decision-making mechanism of agents. A detailed description of how to map the survey data to the decision-making mechanism can be found in [12].

The schedule table contains mobility activities (trips) of the respondents (in the population table). It contains 8 attributes that determine characteristics of trips such as `departure_time`, `distance_of_trip` and `purpose_of_the_trip`. The number of entities in the schedule table depends on the population table. Because the population table has the `number_of_trips` (daily) attribute, which is aggregated to find the length of the schedule table (see Eq. 1).

$$Schedule\ length = \sum_{i=1}^{n} A_{i,number\_of\_trips} \tag{1}$$

Vehicles and resources of the respondents are located in the correspondence table. It has 3 attributes; `ID` (respondent's ID), `type_of_vehicle` and `type of_resource`. The `type_of_vehicle` consists of vehicles (e.g. car, bike, motorbike) with various power-trains (e.g. diesel, gasoline, hybrid, electric). The `type_of_resource` attribute contains mobility resources of agents (e.g. travelcards, driving license, etc.). When trips are performed, the correspondence table is checked to see available mobility modes for a modal choice.

4 Behavior-Driven Demand Model (BedDeM)

BedDeM is an agent-based model (simulation), which aims to capture the heterogeneity of individual demand. In this piece of research, it is calibrated with the mobility data as described above. Thus, it becomes a core tool to model heterogeneous mobility demand in Switzerland [12]. Agents perform their trips based on a decision making mechanism inspired by Triandis Theory of Interpersonal Behaviour (TIB) [12]. The theory involves various psycho-socio and economic determinants such as emotions, habits, monetary cost, social learning, etc. that influence mobility behaviors of agents. After reasoning, agents chose one of the available mobility modes (e.g. car, train, bus, tram, etc.) to perform their trips, called modal-choice.

At the current milestone, we utilize the qualitative data (MTMC and SHEDS) that are described in the previous section. To initialize the model, first, we took a representative sample of respondents from the population table. Individuals in that sample are matched with agents in the model. Basically, each agent employs the real attributes of a respondent. The sampled population consist of 3080 individuals along with 180 attributes. Each individual has a `weight_to_universe` value, which is utilized to fit results to the whole population (i.e. scaling). Some of the attributes are used actively by the simulation whilst some others stay descriptive (transitive). Descriptive attributes are mostly used for the post-processing phase to makes analyses (e.g. describing who are soft-mobility users). Trips of selected respondents are obtained through filtering by respondent IDs. In the same vein, vehicles and resources of the selected respondents are gained (in the correspondence table). Then agents begin to perform their trips in the schedule table simultaneously. They take into account their vehicles such as car and their resources such as public transportation subscriptions in the correspondence table while reasoning. After that, they decide on one of the available mobility modes. Thus, BedDeM generates individual mobility demands of selected respondents (i.e. with which mobility mode they perform their trips). These demands are accumulated to obtain macro-patterns (along with total kilometers) over which the model is calibrated [12]. They should be consistent with real (empirical) data sources. For instance, ca. 22% of trips in Switzerland are performed by train (yearly) [16]. It should be reflected in the macro-patterns that BedDeM generates. Parameters that are used actively in the decision-making mechanism of agents are tuned after several iterations. Detailed information about the calibration process can be found in our recent paper [12].

5 Generation Procedure - Synthesizer

In this section, the method that is followed to generate realistic synthetic data is introduced (see Fig. 1). The idea is generating synthetic data for each mobility profile separately based on narrowed down intra-profile distributions (i.e. smaller attribute spaces) and merging them. In this study, BedDeM is employed with both the real and the synthetic data separately to compare results. This comparison gives an idea about how well the generated synthetic data mimics to the real data. Thanks to the calibration process, macro-patterns that BedDeM generates with the sample data are in line with the macro-patterns in the real data. In other words, accumulated decisions of agents are similar to the respondents'. Hence, comparing simulation results of the real data against the synthetic data illustrates their closeness.

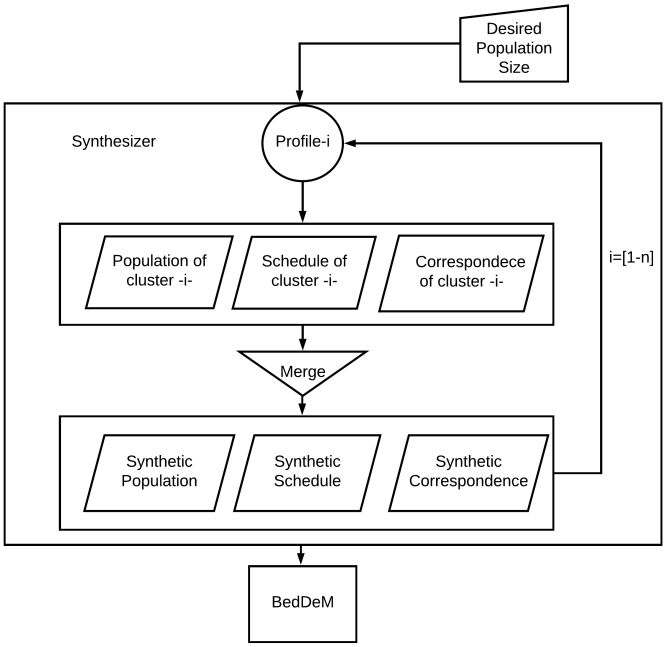

Fig. 1. Architecture of the Synthesizer (n = number of profiles)

We are currently developing a module, called `Synthesizer` (see Fig. 1). It is initialized by the desired population size (the number of agents). According to this input, first, it distributes the entered population size to the mobility profiles according to their sizes, which were obtained in the previous study [3]. Because proportions of the profiles are heterogeneous. Synthetic population, schedule and correspondence tables are generated iteratively for each profile.

Generation of categorical attributes hinges on intra-profile marginal frequencies (distributions) except for location attributes, for which conditional distributions (constraints) are maintained due to legal constraints. For instance, we have two location attributes for the population; `canton` and `type_of_municipality` (according to the definition of the Swiss Statistical Office [16]). Constraints are applied (via cross-tabulations) for them because some municipality types do not exist in some cantons. Therefore, using only marginal distributions might lead to assign some agents in unrealistic places. To generate numeric attributes firstly the type of fitting distribution is detected by statistical tests. For instance the `distance_of_trip` in the schedule table. Firstly, its fitting distribution is detected according to the Cullen and Frey graph (see. Fig. 2).

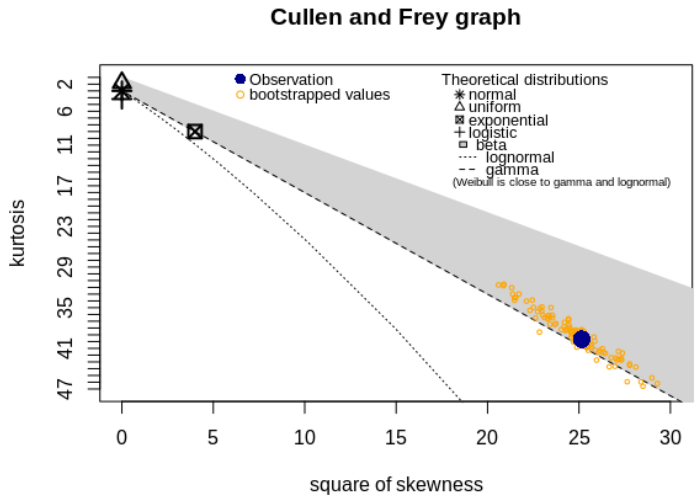

Fig. 2. Illustration of bootstrapped values and fitting distributions

The kurtosis and squared skewness of the real data is plotted along with the bootstrapped values. It seems that possible fitting distributions are the Gamma and the Weibull. After comparing empirical and theoretical values like in Fig. 3, the synthetic values are fitted to the Weibull distribution. This figure compares empirical values against theoretical values in terms of four evaluations; densities, Q-Q (quantile-quantile) plot, cumulative distribution functions (CDFs), and P-P plot (p-value plot). These statistical evaluations give an idea about how fit empirical values to the theoretical ones (i.e. how good is the fitting distribution). Although these three distributions are quite close to each other, the Weibull was better judged by Q-Q plot. These tests are applied for each numeric column. Basically, the fitting distribution of each numeric column is detected and the Synthesizer is configured accordingly. Then, random numeric values within the

boundaries of each profile, which follow the detected fitting distribution, are generated by the Synthesizer. Thus, generated values mimic the same density distribution of the real ones.

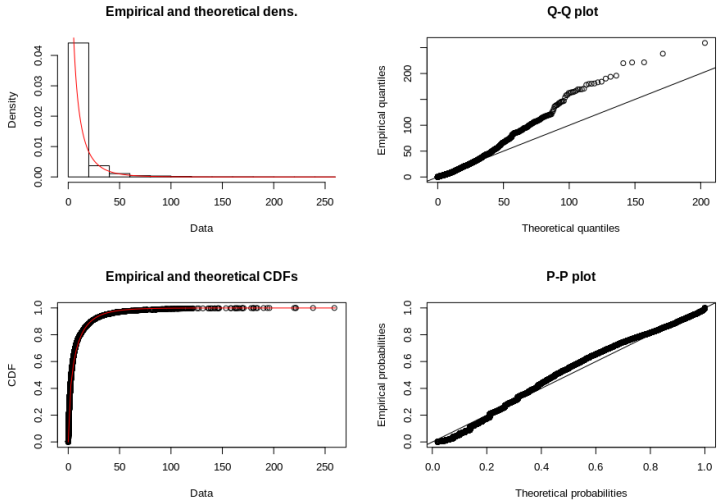

Fig. 3. Comparison of empirical and theoretical values

We obtained 13 mobility profiles and their medoids in the previous study [3]. Basically, respondents in the empirical data were clustered based on their mobility related attributes (i.e. characteristics) to obtain these profiles, which are used for this study as a milestone. The Synthesizer takes the first profile and generates its synthetic data. Then in a loop, all profiles' data is generated successively. Generated synthetic data is merged to obtain the final one. Since attributes (i.e. column names) of generated data of each profile are identical, the merging process (i.e merging synthetic data of each profile) is just aggregating rows to each other. Instead of generation as a whole, the Synthesizer generates data profile by profile (modularity) to benefit less variation (i.e. similar respondents are clustered in the same profile). Once the Synthesizer generates all synthetic data, it re-weights synthetic individuals for scaling. In the MTMC, each respondent has a `weight_to_universe` attribute that indicates how many people are represented by the corresponding individual in the real world. This attribute matters when individual demands are accumulated to obtain macro-patterns. Numbers (e.g. kilometers) of each individual is multiplied by its weight_to_universe value to fit whole Swiss population. Therefore, weight_to_universe values of synthetic individuals should be re-calculated according to the entered population size (i.e. weight_to_universe values are different for 10k and for 100k population sizes due to the difference in resolution). The Synthesizer uses the linear regression as in

Eq. 2 to assigns weight_to_universe values for the synthetic individuals based on the following attributes.

$$\begin{aligned} \text{Weight\_to\_universe}_t = \alpha &+ \beta_1 \text{Canton}_t \\ &+ \beta_2 \text{Municipality\_type}_t \\ &+ \beta_3 \text{Household\_size}_t \\ &+ \beta_4 \text{Income\_level}_t \\ &+ \beta_5 \text{Education\_level}_t + \epsilon \end{aligned} \tag{2}$$

These attributes are selected based on a regression analysis. After new values are assigned to synthetic individuals (i.e. according to the entered population size), generated synthetic data become ready to be used by the simulation.

6 Evaluation

We generate three synthetic populations (along with their schedule and correspondence) with different sizes; 2000, 10000, and 20000 individuals. As has been mentioned, the real data (sample) contains 3080 individuals. BedDeM employs the synthetic and real populations separately (i.e. four different configurations) to check how well the synthetic data mimics the real data (reference). Since The Synthesizer adjusts weight_to_universe values according to population sizes (i.e. to fit whole Swiss population), macro-patterns and total kilometers become comparable. As we can see in Fig. 4, the dissimilarity between the simulation outputs of the real and the synthetic data shrinks with increasing population size (the bars are sorted according to the error rates). It means that the synthetic data become more realistic with increasing population size. Both macro-patterns and total kilometers are in line with the reference.

In Table 1, the absolute differences (errors) between the synthetic and the real populations are illustrated numerically. The error in total kilometers decrease with increasing population size. It shows that the Synthesizer is quite successful to generate synthetic distances for trips (in the schedule table). Both the detection of the fitting distribution and the data generation accordingly seem satisfactory. The error between macro-patterns (mode by mode comparison) looks higher than the total kilometers. The reason might be attribution of vehicles and resources in the synthetic correspondence table.

Table 1. Error rates of different synthetic population sizes

Population Size	Error in Total Kilometers (%)	Error in Macro-patterns (%)
2k	5.6	8.9
10k	0.5	5.4
20k	0.4	5.0

In addition to comparing the simulation results, marginal frequencies of the randomly chosen control attributes are displayed in Table 2. The proportions of

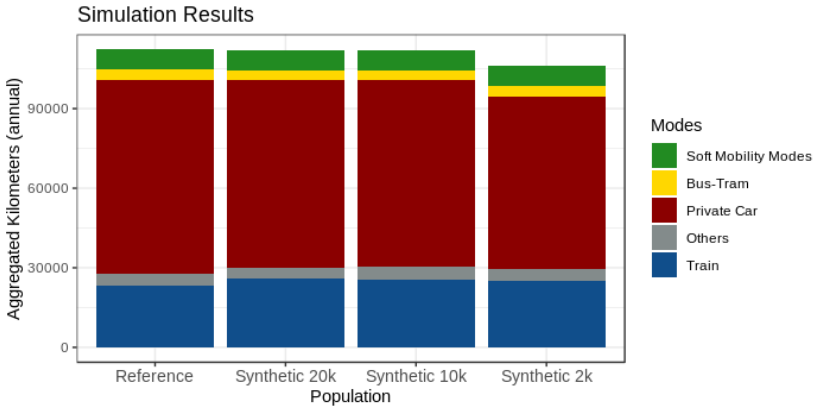

Fig. 4. Outputs of BedDeM with different configurations (The kilometers in the y-axis are million kilometers)

the categories seem quite close. There is no significant difference between there synthetic populations. But some categories with very few proportions, disappear with decreasing population size. The results show that generated synthetic data mimic the real data pleasingly already at this early stage of development. The study is quite open for future extensions that are discussed in the future work section.

Table 2. Comparison of the marginal frequencies (proportions)

Marginal Frequencies of the Control Attributes					
	Categories	2k	10k	20k	Reference
Household_size	1	0.265	0.260	0.259	0.263
	2	0.402	0.416	0.413	0.414
	3	0.139	0.136	0.138	0.137
	4	0.149	0.139	0.142	0.141
	5	0.037	0.040	0.039	0.038
	6	0.005	0.005	0.005	0.005
	7	0.000	0.001	0.003	0.003
Income_level	1	0.019	0.027	0.026	0.022
	2	0.090	0.092	0.093	0.091
	3	0.200	0.195	0.187	0.184
	4	0.301	0.281	0.284	0.274
	5	0.144	0.139	0.143	0.141
	6	0.098	0.101	0.105	0.108
	7	0.057	0.063	0.061	0.061
	8	0.033	0.040	0.038	0.045
	9	0.054	0.058	0.058	0.069
Number_of_cars	0	0.175	0.172	0.168	0.218
	1	0.515	0.505	0.511	0.510
	2	0.268	0.269	0266	0.235
	3	0.038	0.039	0.040	0.337
	4	0.005	0.009	0.008	0.006
	5	0.003	0.001	0.001	0.001

7 Conclusion

In this piece of research, we generate synthetic populations along with mobility activities. Mobility profiles are utilized for data generation. Through profiling (clustering), attribute distributions are narrowed down (i.e. the variation in the clusters is shrunk). We've developed a module called Synthesizer, which generates synthetic data for each mobility profile separately. Then the generated data is merged to obtain the final synthetic data. The Synthesizer utilizes fitting distributions (first detect, then generate) for numeric attributes. For categorical attributes, univariate marginal frequencies (intra-cluster/profile) are employed. 3 synthetic populations with activities are generated with different population sizes. The model employs both the real and the synthetic data for validation. The first results show that increasing population size makes synthetic populations more realistic. Marginal frequencies of control attributes are also checked. Heterogeneity of individuals is maintained. The frequencies are in line with real data. In conclusion, generated data through mobility profiles mimic real data fairly well.

8 Limitations - Future Work

When categorical attributes are generated, only univariate marginal frequencies are considered (except for location attributes). In other words, constraints among attributes are ignored. In the next step of our developments, these constraints can be maintained to improve the results. For numeric attributes, detection of fitting distributions is made manually. Distributions are detected, then the Synthesizer is configured accordingly. In the next steps, it can be fully automatic. The Synthesizer detects the most appropriate fitting distribution automatically based on some statistical tests and generates synthetic data accordingly. Although numerical attributes are generated randomly based on detected fitting distributions, categorical ones are still fitted to marginal frequencies. It is still a kind of cloning. Generating reasonable and realistic white noises around medoids of profiles might help to overcome that problem in the future work.

9 Acknowledgments

This research is part of the activities of SCCER CREST, which is financially supported by the Swiss Commission for Technology and Innovation (Innosuisse). As data sources, Mobility and Transport Microcensus (MTMC) and Swiss Household Energy Demand Survey (SHEDS) are utilized [17, 16].

References

1. Antonini, J.P., Vuidel, G., Klein, O.: Generating a located synthetic population of individuals, households, and dwellings. Tech. rep., LISER (2017)

2. Arentze, T., Timmermans, H.: Albatross: a learning based transportation oriented simulation system. Citeseer (2000)
3. Bektas, A., Schumann, R.: How to optimize gower distance weights for the k-medoids clustering algorithm to obtain mobility profiles of the swiss population, to be published and represented in the Swiss Conference on Data Science (SDS) Conference - June 2019
4. Casati, D., Müller, K., Fourie, P.J., Erath, A., Axhausen, K.W.: Synthetic population generation by combining a hierarchical, simulation-based approach with reweighting by generalized raking. Transportation Research Record **2493**(1), 107–116 (2015)
5. Deming, W.E., Stephan, F.F.: On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. The Annals of Mathematical Statistics **11**(4), 427–444 (1940)
6. Farooq, B., Bierlaire, M., Hurtubia, R., Flötteröd, G.: Simulation based population synthesis. Transportation Research Part B: Methodological **58**, 243–263 (2013)
7. Frick, M.: Generating synthetic populations using ipf and monte carlo techniques: Some new results. [Arbeitsbericht Verkehrs-und Raumplanung] **225** (2004)
8. Harland, K., Heppenstall, A., Smith, D., Birkin, M.H.: Creating realistic synthetic populations at varying spatial scales: A comparative critique of population synthesis techniques. Journal of Artificial Societies and Social Simulation **15**(1) (2012)
9. Huynh, N., Namazi-Rad, M.R., Perez, P., Berryman, M., Chen, Q., Barthelemy, J.: Generating a synthetic population in support of agent-based modeling of transportation in sydney (12 2013). https://doi.org/10.13140/2.1.5100.8968
10. Jain, S., Ronald, N., Winter, S.: Creating a synthetic population: A comparison of tools. In: Proceedings of the 3rd Conference Transportation Reserch Group, Kolkata, India. pp. 17–20 (2015)
11. Jeong, B., Lee, W., Kim, D.S., Shin, H.: Copula-based approach to synthetic population generation. PloS one **11**(8), e0159496 (2016)
12. Nguyen, K., Schumann, R.: On developing a more comprehensive decision-making architecture for empirical social research: Lesson from agent-based simulation of mobility demands in switzerland, to be published and represented in the Multi-Agent-Based Simulation (MABS) workshop - May 2019
13. Pinjari, A.R., Bhat, C.R., et al.: Activity-based travel demand analysis. A Handbook of Transport Economics **10**, 213–248 (2011)
14. Salvini, P., Miller, E.J.: Ilute: An operational prototype of a comprehensive microsimulation model of urban systems. Networks and spatial economics **5**(2), 217–234 (2005)
15. Srinivasan, S., Ma, L.: Synthetic population generation: A heuristic data-fitting approach and validations. In: 12th International Conference on Travel Behaviour Research (IATBR), Jaipur (2009)
16. Swiss Statistical Office (BFS): Mobility and Transport Microcensus (MTMC) (2015), uRL: https://www.are.admin.ch/are/en/home/transport-and-infrastructure/data/mtmc.html. Last visited on 2019/04/21
17. The Competence Center for Research in Energy, Society and Transition - CREST: Swiss Household Energy Demand Survey (SHEDS) (2018), uRL: https://www.sccer-crest.ch/research/swiss-household-energy-demand-survey-sheds/. Last visited on 2019/04/21
18. Williamson, P., Birkin, M., Rees, P.H.: The estimation of population microdata by using data from small area statistics and samples of anonymised records. Environment and Planning A **30**(5), 785–816 (1998)