AN EVALUATION FRAMEWORK FOR RELIABLE POPULATION SYNTHESIS IN AGENT-BASED MOBILITY SIMULATIONS

Federico Bigi, Corresponding Author
Faculty of Science, Technology and Medicine (FSTM)
University of Luxembourg, Esch-Sur-Alzette, Luxembourg, L-4364
Email: federico.bigi@uni.lu

Taha Hossein Rashidi
Civil and Environmental Engineering,
University of New South Wales (UNSW Sydney), Sydney, Australia, NSW 2052
Email: rashidi@unsw.edu.au

Francesco Viti
Faculty of Science, Technology and Medicine (FSTM)
University of Luxembourg, Esch-Sur-Alzette, Luxembourg, L-4364
Email: francesco.viti@uni.lu

Word Count: 7389 words + 0 table(s) × 250 = 7389 words

Submission Date: December 29, 2024

ABSTRACT

This paper presents a comprehensive and innovative evaluation framework for identifying a reliable population synthesis for agent-based modeling - transportation-oriented simulations (ABM - TOS). We show, via this framework and different metrics for the analysis of the generated distribution of the individuals' attributes, that population synthesizers may fail to correctly replicate the real population heterogeneity due to diverse control variables, data limitations, and post-simulation computation of certain parameter distributions. To show these shortcomings, the authors propose a systematic classification of different types of distributions crucial for mobility simulations. The proposed framework aims to provide a comprehensive overview of the population and serve as a rapid 'debugging' tool to identify and rectify any flaws in a specific population during the calibration of the activity-based mobility simulation models. To prove the effectiveness of this framework, we applied it to synthetic populations generated through MOBIUS, a newly developed synthetic population generator, which in this case was employed to create different variants of the Luxembourg population (1%, 10%, 30%). The application of our framework to these populations not only provided an effective method for assessing their goodness of fit, but also helped highlight the distributions that are most critical to the successful implementation of the methodology.

Keywords: Mobility Simulation, Agent-Based Modelling, Population Synthesis, Agent-Based Simulation

INTRODUCTION

Agent-based models (ABM) applied in the transportation domain rely strongly on the correct generation of the travel demand, which is represented by a synthetic population of households and individuals characterized by sociodemographic attributes and daily activity patterns in time and space that at best reflect the real population in an area of study.

The reliability of the synthetic population generation (SPG) process, and in general the level of realism related to ABMs are critical problems in ABM simulations, which have been extensively used in different research domains beyond transportation, ranging from economics, ecology, environmental science, epidemiology, and more recently also in the context of digital twins (*1–4*).

Focusing on ABM for transportation-oriented simulations (ABM-TOS), the population synthesis operation plays an important role, given that it generates the input demand for the simulation environments, a fundamental step for any research investigation. This process usually involves the generation of individual-level data that accurately reflects the demographic and socio-economic characteristics of a given population, as well as information regarding their travel mobility patterns. These synthetic populations are then used to simulate various aspects of urban mobility, including travel behavior and transportation infrastructure use.

Synthetic population generation processes can generally be divided into two stages: *fitting* and *generation* (*5, 6*). While the fitting stage deals with the adjustment of an initial distribution to match known marginal distributions, for example, the different households per area, the generation stage involves creating synthetic individuals from these distributions (*7*).

Over the years, a multitude of methodologies and techniques have been developed to tackle the challenges associated with population synthesis. There are usually two types of data sources that can be gathered for generating a synthesized population, either from publicly available data or from travel surveys (*6*). Usually, these sources are associated with spatial zoning systems and can be available in two forms: individual agent samples and cross-classification tables. Zoning systems, defined based on factors such as maximum density levels and physical barriers, often provide some level aggregations, which can range from regional to communal. On the other hand, agent samples, such as Public Use Micro Sample (PUMS) in North America or Sample of Anonymised Records (SARs) in the UK, contain demographic and socioeconomic information but are usually available only for large spatial areas to protect privacy. Cross-classification tables, released by statistics bureaus at different zoning levels, provide socioeconomic and demographic data in 1-3 dimensional tables (*6*).

Achieving a good level of accuracy in reproducing the population characteristics heterogeneity presents several challenges. One of the primary obstacles is the unavailability of complete data for the population, often due to privacy concerns or logistical constraints. This issue might be further worsened by the use of proprietary software, which may limit accessibility and transparency in the data synthesis process. This, paired with the lack of high-resolution geo-location information in available datasets can lead to inaccurate models and simulate spatially-dependent phenomena. To overcome this problem, multiple techniques have been applied to this topic to face the multiple problems linked to the scarcity of data. In particular, for the fitting stage, we can identify in the literature three macro-areas: the data scaling techniques, through methods such as the Iterative Proportional Fitting (IPF); the distribution extraction, with techniques such as the Markov Chain Monte Carlo (MCMC) methods, up to more advanced different Machine Learning (ML) models.

No matter the methodology or sources, what the synthesizers have to deal with is fundamentally, in the fitting stage, the extraction of the overall underlying picture from the available data, typically in the form of different distributions, and the generation of the correct population in the *generation* stage. The fact that population synthesis typically is prior to the ABM and the related simulations further underlines the importance of ensuring a correct population synthesis process, as errors that are accumulated at this step will propagate in the model system and will inevitably affect the demand forecasts, analysis (*8*). These problems usually depend either on the methodology used for the synthesis of the population, or the absence of data for specific information. Nonetheless, ensuring that the population generated matches, as much as possible, the initial data can ensure a flawless transition toward the calibration step.

With the purpose of developing a systematic assessment of population synthesizers for agent-based mobility models, this paper proposes a new comprehensive framework for the analysis and quantification of the reliability of synthesized population generation processes. This framework is designed to cover and analyze all the different distributions that are important for any ABM-TOS simulation. The methodology will serve two main purposes: firstly, it provides an in-depth overview of the population characteristics highlighting, based on the specific distributions, where the synthetic data are actually reflecting the extracted initial data; moreover, it functions as a quick 'debugging' tool, enabling quick identification and correction any inaccuracies in a population synthesis process.

The remainder of this paper is organized as follows. In Section 3, we will go through the different methodologies developed for the population synthesis, highlighting their strengths and weaknesses, and underlining the absence of a no generalized comprehensive framework for analyzing the goodness of fit of a synthetic population for ABM-TOS with respect to the initial data. Then, in Section 4, we will explain each category developed in the framework. Then, we present the case for MOBIUS, a newly developed population synthesizer in Section 5, where a population synthesizer has been developed for the MATSim Luxembourg Scenario. In Section 6, the results of the application of the framework to the 3 synthesized populations for Luxembourg are presented, together with the detailed categorization choice, which is then discussed in Section 7, Finally, Section 8 concludes the paper and presents the next step of the research.

LITERATURE REVIEW

In this section, we present the literature regarding different methodologies and techniques widely used in population synthesis for different ABM simulations, namely Iterative Proportional Fitting (IPF) approach, Markov Chain Monte Carlo (MCMC) methods, and the use of Machine Learning (ML) models, together with some of the different population synthetizers that uses the specific technique. However, what emerges from our review is the lack of a structured and comprehensive methodology for assessing the performance of population synthesis. This gap in the literature motivates our proposal for a new framework designed to analyze all the different distributions that are crucial for mobility simulations.

Iterative Proportional Fitting

The Iterative Proportional Fitting (IPF) has been first described by Deming et al. (*9*) and is also known as matrix ranking, RAS method, or matrix scaling. It is a mathematical method used in statistics to adjust the values of an initial matrix in such a way that its marginal totals align with independent estimates while preserving the relative proportions within the matrix. Important features of IPF are the minimization of relative entropy and preservation of cross-product ratios; in other words, among all contingency tables that satisfy the marginal constraints, the resulting table is the most similar to the initial table. The strength of IPF lies in its ability to preserve the relationships between different attributes in the data while ensuring that the synthetic population aligns with known marginal distributions.

While the IPF is effective in different contexts, it also comes with certain limitations. One such issue is the "zero-cell problem", which arises when the initial seed matrix contains cells with a value of zero, due to for example the absence of observations in a specific area in the context of the synthetic population. Since IPF operates by adjusting cell values proportionally and iteratively, a cell with a zero value will remain zero throughout the adjustment process, regardless of whether the marginal totals suggest that the cell should have a positive value. In this context, this can result in a population that does not fully represent the diversity of the actual population, particularly in terms of less common demographic or socio-economic groups. In the context of different Population Synthesizers for ABM simulations, IPF has been used to adjust initial multi-dimensional distributions to match known marginal distributions, such that each individual in the population sample is assigned a weight such that the weighted population shows predefined marginal distributions for example for age, gender, and other attributes or combinations thereof (*10, 11*).

Beckman et al. (*12*) present a methodology for generating synthetic populations based on census data, proposing an algorithm that uses the IPF to estimate the proportion of households in a block group or census tract with a desired combination of demographics. The synthetic population is then created by selecting households from the Public Use Microdata Sample (PUMS) in proportion to the estimated entries. Given that no ABM simulation is performed, the metrics analyzed to estimate the closeness of the generated data to the reference data are both household size distributions, as well as multi-variate distributions like the age distribution per gender. Horl et al. (*1*) present a methodology for generating synthetic travel demand using open and publicly available data. The process utilizes the PopGen package and OpenStreetMap data to provide location candidates. Agriesti et al. (*13*) presents a methodology for assigning a synthetic population using simPop (*14*) for ABM-TOS assignment using available public data, paired with an anchor points approach for their design. The methodology is implemented and validated in a case study for the city of Tallinn, Estonia. In this work, given that it was focused mostly on the ABM assignment, they evaluate more "zonal" parameters, such as the home-primary activity location distribution, the spatial assignment of the workplaces as well as the spatial distribution of the households.

Jain et al. (*15*), presents a comparative study of two synthetic population generation tools, PopSynWin and PopGen to create synthetic populations that closely represent the actual demographics of the study area, scaling them up using the IPF. The generated synthetic populations were then validated against actual aggregate census data to evaluate their representativeness through the analysis of the different sociodemographic parameters (gender, age, employment), as well as the household size Finally, Tozluoglu et al. (*16*) presents the documentation for their population synthesizer, Synthetic Sweden Mobility (SySMo). In this case, they use classification for the different attributes for the population to synthesize: the *basic attributes*, defined as the different sociodemographic characteristics of an agent (gender, age, households); the *advanced attributes*, defined as the employment and student status, the income and the car ownership. They integrated the activity, the location, and the mode assignment in the synthesis process, using the IPF in the primary activity location assignment, paired with a gravity model, assessing the goodness of fit of the generated population by looking at different parameters, such as the basic variables (gender, age, households), some advanced ones such as gender per age, measures of employment, household size and so on.

MCMC

At the other end of the spectrum, Markov Chain Monte Carlo (MCMC) methods are a class of algorithms used in computational statistics for sampling from a probability distribution (*17*). MCMC methods work by constructing a Markov chain that has the desired distribution as its equilibrium to the reference distribution. The state of the generated chain, after a specific number, is then used as a sample of the desired distribution. The quality of the sample improves as a function of the number of steps.

One of the advantages of MCMC methods over the IPF is that they can handle more complex structures and dependencies in the data while being also less prone to problems like the zero-cell issue that can occur with IPF. On the other hand, MCMC methods can be computationally intensive, particularly for large data sets or complex models. As for the different population synthesizers, Felbermair et al. (*18*) present a methodology for generating a synthetic population with activity chains for ABM-TOS, using both Bayesian networks and Markov Chain Monte Carlo methods to synthesize a population with activity plans based on limited survey data. The simulation results in a good adherence with the initial distributions of employment rates, as well as trip-length distributions.

Farooq et al. (*6*) propose a Markov Chain Monte Carlo (MCMC) simulation-based approach for synthesizing populations for use in urban systems evolution microsimulations. Their methodology better captures the heterogeneity of populations, handling both discrete and continuous attributes, scaling more efficiently as the attribute grows, and showing the effectiveness of the proposed method by comparing it with standard IPF using real population data from the Swiss census and synthesizing a population for Brussels, Belgium, where data availability was limited. Sun et al. (*19*) propose a novel method for population synthesis using Bayesian networks to reproduce the joint probability distribution of the studied population system. introducing also the concept of the Markov blanket, which allows for efficient inference and sampling of a node by considering only its immediate connections, rather than the entire network. The approach is then validated using the Household Interview Travel Survey (HITS) data for Singapore.

ML Models

Machine Learning (ML) models are increasingly being utilized in population synthesis processes, proposing interesting methodologies given their capability of faithfully reproducing complicated distributions underlying the input data. These models leverage advanced algorithms to address issues such as data scarcity and privacy concerns while generating synthetic populations that closely mirror real-world demographics and behaviors.

Arkangil et al. (*20*) proposes a framework to generate a synthetic population that includes both socioeconomic features (e.g., age, sex, industry) and trip chains (i.e., activity locations), showing that the synthetic population generated by the framework closely matches the real population in terms of both socioeconomic features and trip chains.

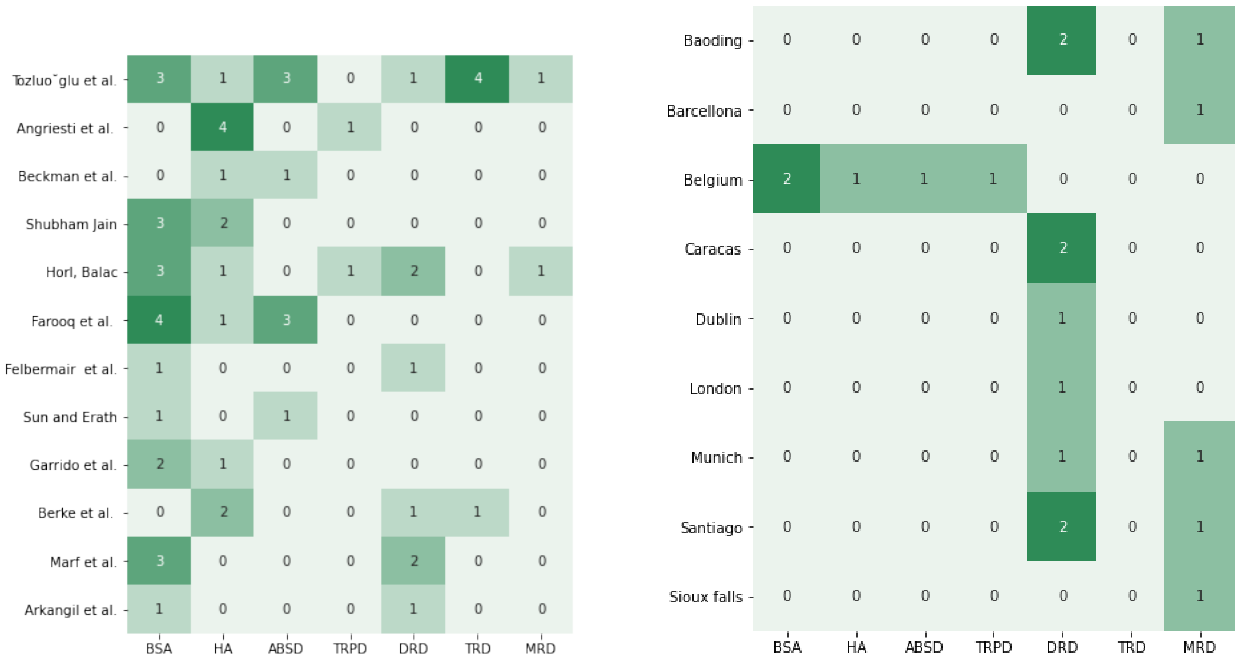
Berke et al. (*21*) present a framework for generating synthetic mobility data using a deep recurrent neural network (RNN) trained on real location data. Badu-Marfo et al. (*22*) use a deep generative model called Composite Travel Generative Adversarial Network (CTGAN) for synthesizing population data in agent-based transportation modeling, reconstructing composite synthetic agents with tabular and sequential mobility data. The results showed that CTGAN outperforms other models in terms of data utility and privacy. Garrido et al. (*8*) propose the use of deep generative latent models, specifically the Wasserstein Generative Adversarial Network (WGAN) and the Variational Autoencoder (VAE) to generate agents for agent-based modeling.

These models are capable of learning a compressed representation of the data space, effectively imputing missing information when projected back to the original space. The validation performed through a Danish travel survey with a feature-space of over 60 variables demonstrates their effectiveness in generating diverse yet valid synthetic attribute combinations.

Research Gap and proposed approach

The issue with the population synthesizers provided in the literature, given their diverse objectives and the different input data, is that they do not all offer a structured analysis for assessing the quality of the generated population for use in ABM-TOS. To systematically analyze these existing methods, we propose a framework that classifies different types of distributions that are important for any ABM-TOS. This framework, which we will elaborate more in detail on in Section 4, includes the following categories: Basic Sociodemographic Attributes (BSA), Household Attributes (HA), Advanced Sociodemographic Distributions (ASD), Tripchain Related Distributions (TRPD), Distance Related Distributions (DRD), Time-Related Distributions (TRD) and Mode Related Distributions (MRD). By applying this framework to the existing literature, we aim to highlight the gaps and inconsistencies in current methods. To demonstrate the effectiveness of this analysis, we also analyzed the majority of MATSim scenarios presented in (*23*), where 9 scenarios were selected, providing results that can be classified through our framework. The analysis is presented in Figure 1.

From our analysis of the literature (Figure 1a) and existing MATSim scenarios (Figure 1b), it becomes clear that many validation analyses fail to address the aspects that are important to achieve a reliable ABM; in this case, we decided to showcase on the heatmaps only the absence/presence of a specific analysis given a specific population, while not focusing directly on the emphasis that a specific population synthesizer has in terms of a given group of distributions. Regarding Figure 1a, as we can see most of the literature focuses on the BSA and HA as part of their validation process, while failing to address metrics such as the TRD and MRD. This can be justified though through several factors: the diverse objectives served by different population synthesizers, which may not be solely focused on serving an ABM-TOS; the potential absence or non-inclusion of certain data in the generated population; and the possibility that some distributions are only computed post-simulation, rather than beforehand. Looking at the scenarios analyzed for the ABM-TOS instead (Figure 1b), we can see as almost none of them manages to cover all the related aspects that impact the different mobility simulations. While at least the Belgium scenario covers the majority of the population-related aspects, it fails to address the ones that are more regarding the mobility-related distributions (DRD, TRD, MRD). On the other hand, while it is understandable that all the other analyzed scenarios might want to present more analysis on mobility-related distributions, they completely fail to cover both the TRD and TRPD distributions, which are essential to guarantee the correctness of the analysis in the ABM-TOS. The lack of a comprehensive, structured analysis for assessing the quality of a generated population for ABM simulations can lead to overestimations or underestimations of some population characteristics or behavioral aspects, as confidence might be placed in other analyzed parameters (for ABM-TOS, these might be seen as for the modal split, trip length). To address this, we propose a framework



(a) (b)

FIGURE 1: (a) Literature validation results classified; (b) MATSim scenarios validation results classified.

designed to analyze a large set of different distributions that are crucial for mobility simulations. This framework will serve a dual purpose: firstly, it will provide a comprehensive overview of the population, ensuring that our generated data accurately reflects the underlying reality; secondly, it will offer a rapid 'debugging' tool, making it easier to identify and rectify any flaws in a specific population during the calibration of the ABMS models.

METHODOLOGY

For our methodological framework, we decided to group the different parameters usually analyzed into different categories, together with new distributions extracted from the literature, that proved to have an impact on the ABM and ABM-TOS.

Basic Sociodemografic Attributes (BSA)

Hanson et al. (*24*) define the sociodemographics for travel-activity patterns as both the socio-economic and role-related characteristics of the individual, measuring the occupation, level of education, income, gender, marital status, and stage in the life cycle. These characteristics have proved to be one of the determinants for different aspects of traveling (e.g., frequency of travel, dispersion of destinations visited) (*25*). In this case, what should be included in this category would be the fundamental sociodemographic parameters. These are including, but not limited, to gender distribution, age, employment status, income distribution, and driving license ownership.

Household Attributes (HA)

Bradley et al. (*26*) show that intra-household interactions and characteristics constitute an important aspect in modeling both activity and travel-related decisions, influencing also decisions such as daily activity pattern (DAP) of the single participants in the household. Hu et al. (*27*) shows that there is a direct correlation between the travel and activity behavior of a household and its specific attributes, such as its location, car ownership allocation, and primary activity allocation. In this category, what we can expect to find are, and again not limited to, distributions of the size of the household, household location, number of cars owned by the household.

Advanced Sociodemographic distributions (ASBD)

In this category, the idea is to provide some bivariate distributions, given that usually these fall out of the control variable area. This can help us understand if the methodology underlying the creation of both sociodemographics and household collides together in a correct way. The bivariate distributions that can be analyzed here can be therefore a combination of what we defined as the Basic Sociodemographic Attributes and Household Attributes. Examples include age-household size, age-gender, and gender-household size distributions, as well as the percentage of participation in primary/secondary activities, home to primary activity distributions (work, school).

Trip-chain related distributions (TRPD)

In ABMS, as well as in activity-based modeling, trip chain assignment plays a critical role in determining the sequence of activities and travel events for each individual or household, and therefore the behavior of the generated population. This sequence, also known as a trip chain, includes details about the activity type, activity location, timing, and duration of each activity, as well as the mode of transportation used to travel between activities. Scheffer et al. (*28*) shows that trip chain assignment can not only vary from a temporal point of view, but especially from a modal split point of view, affecting the different choices that then are simulated through the ABMS. They show that the importance of modeling dynamic mode choice with trip chaining and time of the day affected, in their case study, the modal share of car drivers by more than 40% between hours of the day and about 30% between different activities. In this category, what we can expect to find are, and again not limited to, trip chain distribution over the populations, and primary-secondary activity location patterns.

Distance-related distributions (DRD)

The correct computation of distance distributions in any Activity-Based Model is of fundamental aspect for achieving accurate ABMS and understanding the travel behavior of the generated population. Distance, as a key determinant of travel cost and time, directly influences individuals' decisions about when, where, and how to conduct their activities, as well as the choice of mode of transportation, the timing of activities, and the overall pattern of daily travel. For instance, longer distances might encourage the use of faster but potentially more expensive modes of transport, while shorter distances might make walking or cycling more attractive options. Examples of attributes that could be found in this category include estimation of total distance traveled, home-primary activity distance distribution, and trip length distribution.

Time-related distributions (TRD)

The assignment of activity durations directly influences the temporal structure of the synthetic population's daily activities and is influenced by the distances to be covered, as well as the trip chain assigned. The duration of an activity can affect the choice of mode of transportation, the route chosen, and the timing of subsequent activities, playing also a significant role in determining the level of congestion in transportation networks at different times of the day. Moreover, accurately modeling activity durations can provide valuable insights into the population's time use patterns, which are essential for understanding and predicting behavioral responses to policy changes or disruptions. For this category, examples of attributes that might be included are the duration assigned for primary and secondary activities and activity start times.

Mode-related distributions (MRD)

Mode-related distributions play a pivotal role in Activity-Based Modeling (ABM), as they capture the diversity of transportation options available and their usage patterns among the population. These distributions reflect the choices individuals make regarding their preferred mode of transport, which can be influenced by a variety of factors such as distance, cost, convenience, and personal preferences. In the context of ABM, an accurate representation of mode-related distributions is essential for simulating realistic travel behavior. For instance, the choice of transportation mode can significantly impact the duration of activities, the sequence of activities (trip-chaining), and the distances traveled. Examples of distributions that might be included in this category are the total distance traveled per mode and the modal split.

Metrics analysed

To quantify the reliability of each parameter, we need to use metrics that are capable of assessing the similarity of different variable distributions. For our specific analysis, the Hellinger distance, the NRMSE and the JS divergence were chosen to compare the generated distributions with the reference ones.

The Hellinger distance is commonly used in statistics and information theory to measure the similarity between two probability distributions. It measures the similarity between the square roots of the values in the two distributions, computed as follows:

$$H(P,Q) = \sqrt{\frac{1}{2}\sum_{i=1}^{n}(\sqrt{p_i} - \sqrt{q_i})^2} \tag{1}$$

where P and Q are the two probability distributions being compared, n is the number of values in the distributions, and p_i and q_i are the values at the ith index of P and Q, respectively. This metric was chosen mostly because of its sensitivity to variations in the shape of the distributions.

Furthermore, the Normalized Root Mean Square Error (NRMSE) was chosen as a second metric to evaluate the difference in magnitude of the selected distributions. The NRMSE is a measure used for assessing the similarity between two probability distributions, and is defined as follows:

$$NRMSE = \frac{1}{N}\sqrt{\sum_{i=1}^{N}\left(\frac{y_i - \hat{y}_i}{y_{max} - y_{min}}\right)^2} \tag{2}$$

where y_i and $\hat{y}_i$ represent the observed and estimated values, respectively, y_{max} and y_{min} are the maximum and minimum observed values, and N is the total number of observations.

Furthermore, to complement our analysis with a metric to evaluate the goodness of fit of each distribution, we chose to add to our evaluation methodology the Jensen-Shannon (JS) divergence. This symmetric method measures the similarity between two probability distributions. The JS divergence is defined as the mean of the Kullback-Leibler (KL) divergence of P from the average distribution M, and the KL divergence of Q from M, where P and Q are the two probability distributions in comparison.

$$JSD(P,Q) = \frac{1}{2}D_{KL}(P||M) + \frac{1}{2}D_{KL}(Q||M) \tag{3}$$

In this equation, D_{KL} is the Kullback-Leibler divergence and M is defined as the average of P and Q: $M = \frac{1}{2}(P+Q)$. Therefore, the JS divergence calculates the divergence of P and Q from the average distribution.

With these three metrics, now we can have a clear overview on the difference in terms of magnitude, shape and goodness of fit, hence providing a more robust evaluation tool for the analysis of the congruence of the different generated distributions.

Finally, it is worth noting that the choice of these measures is also driven by the qualitative perspective this framework seeks to provide. The target of this framework is nonetheless to provide a deep and quick understanding of all the fundamental distributions that affect the population synthesis for ABM-TOS, as well as the differences between the reference and generated distributions, in a way that every discrepancy can then be further investigated through proper quantitative analysis.

CASE STUDY

In this section, we discuss and describe the process underlying the generation of the synthetic population by instantiating the process using the data from a large travel survey conducted in Luxembourg in 2017. The Luxmobil Survey 2017 was conducted by the Ministry of Mobility and Public Works and involved 40,000 residents (out of a total of about 600,000 inhabitants in 2017, hence nearly 7% of the total country's population) and 45,000 cross-border workers (out of around 200,000 daily commuters, hence around 22%). The response rates were high enough (around 30%) to make the results representative and unbiased, obtaining around valid 35000 answers, with a balanced distribution of the answers with respect to the actual population (73% Luxembourgish, 12% French, 7% Belgian, and 8% German respondents). After cleaning the data from e.g. people reporting incomplete or illogical trip chains, around 22000 valid responses were retained. The cleaned dataset represents around 2,8% of the daily travelers in the country, which was considered enough to start the procedure for the population synthesizer. Additionally, the travel survey itself provided a pre-existing zoning structure at administrative unit level, which is composed of 150 distinct zones.

The individuals responding to the questionnaires reported a day of trips to reach all daily activities and the used mode(s) of transport in the form of travel diaries, in addition to basic sociodemographic information and other questions. From the travel survey, we extracted the following data for this analysis:

- spatial distribution with zonal information, with related facilities location file;

- zonal household and population distribution categorized by the considered attributes;
- zonal trip chain distribution;
- OD matrices per hour, which we computed from the Luxmobil travel survey and validated through PTV Visum;
- zonal distribution of typical activity time duration and departure time for the first activity;

These data were then prepared for a novel population synthesizer, coined MOBIUS (Mobility Optimization Based on Iterative User Synthesis). The macrostructure of MOBIUS has been inspired by two state-of-the-art synthesizers presented in (*16*) and (*1*), respectively. We followed the same general steps presented in the two papers in our approach, namely: Population Synthesis, Activity Generation, and Location and Mode assignment.

In the Population Synthesis phase, data is gathered from the source and scaled to match the target population. Information was collected for each region from the travel survey, including the relative marginals for each specific area (Luxembourg, Germany, Belgium, France). Unlike the synthesizers presented in (*16*) and (*1*), which sample the population post-IPF and then allocate them to households, we chose to follow a reverse approach, i.e. we first created all the households along with their associated attributes such as the household's zone, car ownership, and household composition size. This approach was adopted to facilitate attribute inheritance, which assists in both control and code efficiency. Next, for every household that was created, we selected agents to be generated from the population dataset corresponding to the household's area. These agents were categorized into *adults* ($\geq$ 18 years) and *children* ($<$ 18 years), following the composition of the household, to gain enhanced control over the attribute distribution. This division was done in order to avoid the creation of agents with unrealistic characteristics, such as a 5-year-old worker with a driving license. At this point, the agents' attributes are chosen from the sociodemographic attributes with respect to their age group. Moreover, for the household latitude/longitude pair, given an initial area, a bounding box approach with random sampling was applied.

In the Activity Generation phase, which includes the creation and assignment of trip chains to each agent, we sampled a Tripchain T_i for each agent a_i produced in the Population Synthesis phase. This was based on the initial departing zone and was assigned to the chosen agent. In our specific scenario, we extracted the top 25 trip chains to eliminate outliers, which accounted for 90% of the total trips recorded in the Luxmobil travel survey. If a *children* is chosen for the assignment, we opted to limit the sampling to a non-working activity trip chain.

In the Location Assignment phase, we, following the two above-mentioned state-of-the-art synthesizers, adopted an OD probability matrix sampling approach, segmented in our case by hour and activity type. The Home location assignment, performed during the Population Synthesis step in our case, utilized a bounding box approach on the selected initial zone due to our zoning's granularity. For the departure time of the first activity, we calculated the average and standard deviation of the departure time from each zone for a specific activity. Next, we estimated the next zone based on the current zone and the next activity through an OD matrix per hour and per activity sampling. This allowed us to access the related OD matrix for that specific time frame and sample the destination zone from the corresponding distribution, given the initial zone assigned in the Population Synthesis step. Here, some remarks have to be made: having a relatively detailed zoning for such a small country (2586 km^2 partitioned in 150 zones, resulting in 17.24 km^2 per zone on average) allowed us to further detail the latitude/longitude assignment, instead of using a random sampling bounding boxing approach. In this case, for all the activities that we categorized (work, school, leisure, others) and for each zone, the top 10 lat/long pairs were selected, and during the assignment of the exact location, another sample was performed, on the specific area, to assign the precise destination. In case of a missing activity in a zone, the closest area was chosen to perform the random assignment. For the activity duration assignment, we estimated it from the distribution of that activity's duration per zone on the travel survey.

For the assignment of leg mode, an initial estimation is made based on the beeline distance to be traveled between the center of the two zones and the activity to be performed. However, this is subsequently fine-tuned to align with mobility data through a MATSim simulation. This initial estimation is important despite its inability to account for traffic propagation, as it helps in correctly calculating the end-time of each activity and in turn the activity duration. This is critical to prevent sampling from an inconsistent aggregated OD matrix.

The full assignment loop is detailed in Algorithm 1, which outlines the Synthetic Population Generation Assignment Loop.

Algorithm 1 MOBIUS Assignment Loop

```
for each agent $a_i$ to generate do
    Initialize $T_i$, tripchain of $a_i$, and $time_{curr}$, current time
    for each $act \in T_i$, with $act$ as single activity in the Tripchain do
        if $act$ is the first activity then
            Assign $T_{dep}$, departure time from home, set it as $time_{curr}$
            Estimate Mode of transport and travel time to get $tt_{mode}$
            Update $time_{curr} += tt_{mode}$
        else
            Assign location for $act$ through the OD matrix sampling, let the location be $z_{dest}$
            Sample $t_{act}$, activity duration from distribution of duration of activity in $z_{dest}$
            Estimate Mode of transport and travel time to get $tt_{mode}$
            Update $time_{curr} += tt_{mode} + t_{act}$
        end if
    end for
end for
Return $a_i$ with $act \in T_i$ assigned with times and locations.
```

RESULTS

To showcase an application of the assessment framework on the MOBIUS synthesizer, we proceeded as follows. We first generated three distinct synthetic populations, specifically the 1% ($\approx 6,454$ individuals), 10% ($\approx 64,539$ individuals), and 30% ($\approx 193,617$ individuals) over the total population of $\approx 645,390$ people (as for Luxembourg population of 2021), including both resident and cross-border information, as reflected in the travel survey.

For each of these population shares, we then conducted a MATSim simulation until equilibrium was reached, which was determined by matching the modal split and running the simulation for 150 iterations. After achieving equilibrium, we performed a further calibration using Cadyts (*29*). This calibration was conducted by running a MATSim simulation at 300 iterations, using the output of the previous step, and was based on data from 21 traffic counts recorded in October 2021, which covered around 30% of the total traffic counts presented in our dataset. Road capacities were properly down-sized to match the observed volume-to-capacity ratios.

The reason behind the choice of these 3 population sizes is that, as one might expect, in some distributions we expect the metrics to converge as the population size grow. This is, for example, the case of age distributions, whose error should decrease with respect to the reference data as the sample grows in size. On the other hand, if the error does not decrease as the sample grows, that is where our analysis has to be conducted, as showcased in Section 7 for our case study.

The results of this analysis, applied to the above-mentioned populations and adopting the three metrics described in the previous section, are summarized in Figure 2.

Basic Socio-demographic Attributes (BSA)

The BSA presented in this study includes data regarding error in between the distribution of age and gender. Unfortunately, we could not gather data for the income from the travel survey and is therefore excluded from this analysis.

Household Attributes (HA)

The HA in this study focuses on the distribution of household attributes across different zones. It includes the number of cars assigned (*car_numbers*) and the distribution of family size.

Activity-Based Socio-demographic Distribution (ABSD)

The ABSD in this study includes the OD pair assignment of workers (*workers_arrival_zone*) and students (*students_arrival_zone*) for their primary activity segment, defined as home-work/school, matched at the zonal level.

Tripchain Related Distributions (TRPD)

The TRPD in this study analyzes specifically the trip chain assignment for adult cross-border commuters (*trp_>18_rsd*) and for adult residents (*trp_>18_crb*). This specific cut on the distributions analyzed was chosen due to the specific case and the high impact that cross-border commuters have on Luxembourg's daily traffic situation.

Distance Related Distributions (DRD)

The DRD in this study presents the beeline school (*school_distance*) and work (*work_distance*) distance distribution. In this case, the beeline distance was chosen as a metric given that, opting for a direct approach, should have included the routing with the actual traffic propagation of the

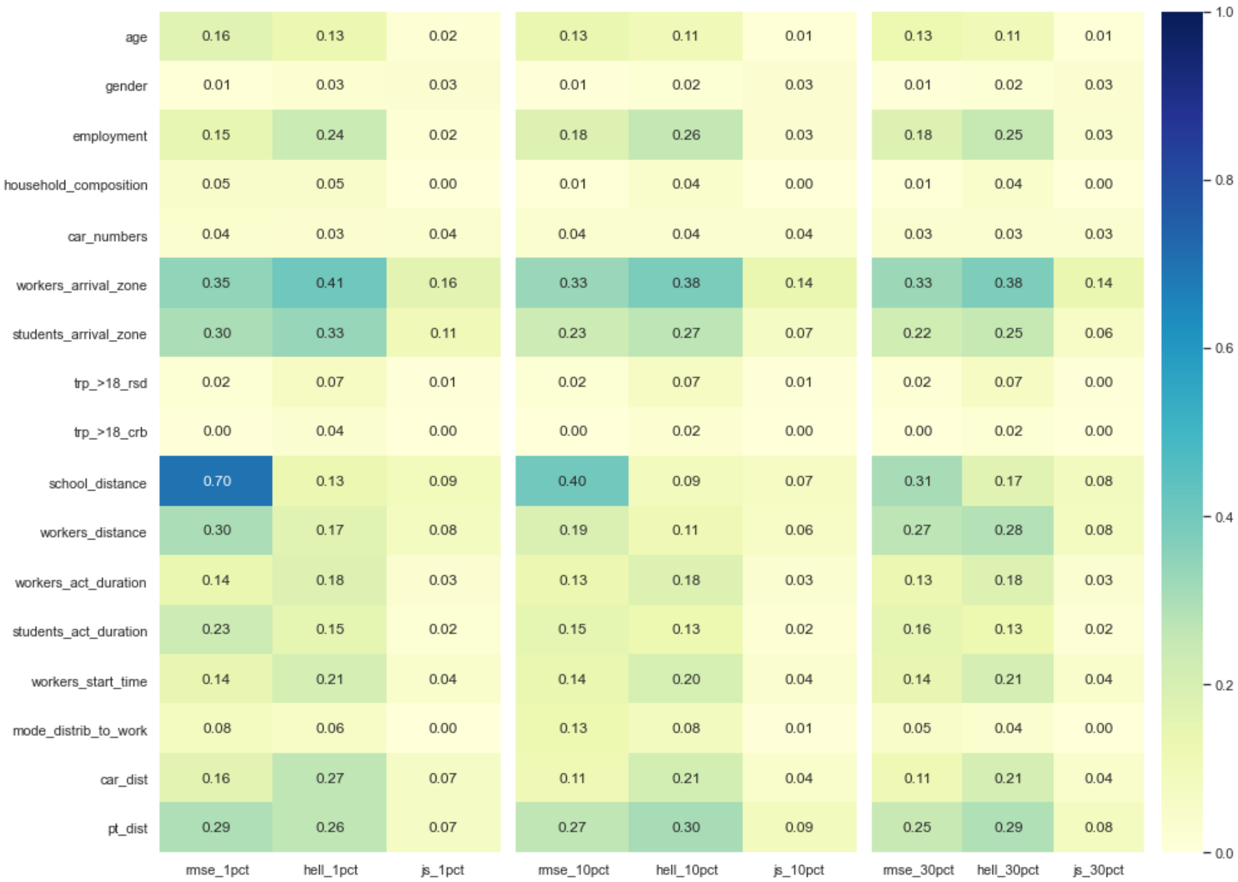


FIGURE 2: Results for the 1%,10% and 30% Luxembourg generated population.

network, while with this choice our aim was to evaluate the effectiveness of the assignment of activity destinations through the OD-pair sampling.

Time-Related Distributions (TRD)

The TRD in this study focuses on the primary activity duration of worker (*workers_act_duration*) and student (*student_act_duration*) distribution, together with the departure time from home towards the primary activity for workers (*workers_start_time*).

Mode Related Distributions (MDR)

The MDR in this study presents the distribution for the primary activity (*mode_distrib_to_work*) and the relative car (*car_dist*) and pt (*pt_dist*) distance distributions. The choice to analyze car and public transport was done given that, for our MATSim simulations, these are the only two modes that we simulated with actual routing and physical propagation, since walking and cycling are assigned by teleportation in the software.

DISCUSSION

Generally, the trend observed in the proposed measures for most of the categories is either stable or decreasing as the synthesized population increases. This is aligned with the expectations and shows the sensitivity and properties of the metrics to identify issues in the generated distributions from the selected sample sizes.

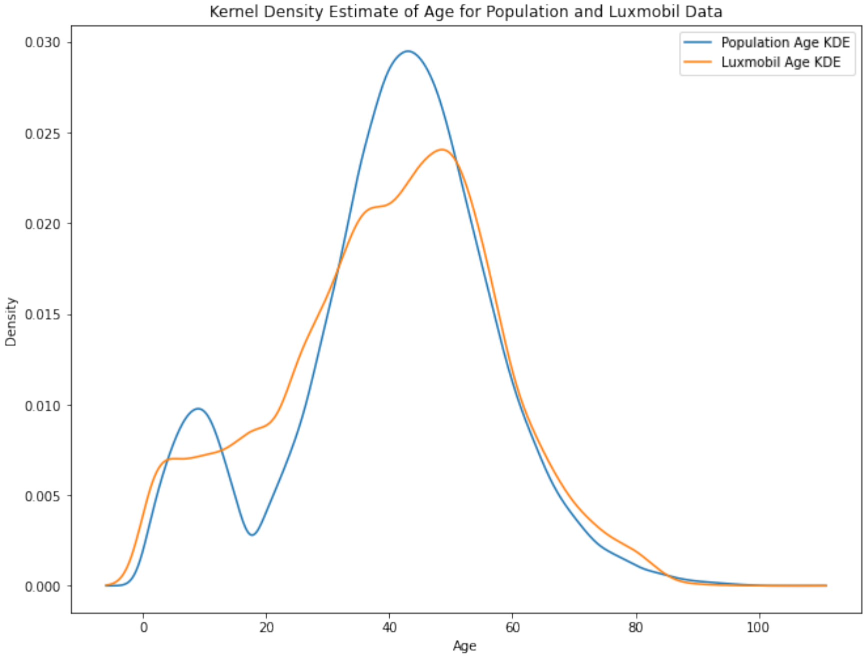


FIGURE 3: Age distribution of 30% of Luxembourg population generated through MOBIUS with the reference distribution.

The decreasing NRMSE and JS, as well as the stable Hellinger distance for the age category, as seen in Figure 2, can be explained by the fact that even if the distributions align more and more between the reference and generated as the population size increases, the shape of the distribution is different. This can be attributed to how we assigned the age of the population versus the reference one, as explained in Section 5, where for the case of MOBIUS a sharper bimodal normal distribution is generated than what the reference data shows. This can be clearly seen in Figure 3.

The metrics generally match our expectations, especially given the high number of controlled variables used in MOBIUS for the BSA and HA. Firstly, the NRMSE for the *school_distance* metrics decreases as the population size increases, as well as the *school_arrival_zone*. This can be explained by the limited number of accessible schools in our network. With a smaller student population, a larger discrepancy is expected, which diminishes as the population size increases. On the other hand, the *worker_distance* and *worker_arrival_zone* metrics do not show significant improvement with an increase in population size, but instead a stable trend. Especially regarding the *worker_arrival_zone*, we can see that the JS metric shows a slight misalignment between the generated and reference data. This could be attributed to one of the core aspects of MOBIUS, the random sampling of the OD segmentation, where the departure time from home to work determines the OD matrix sampled. If the departure time from home is not accurately aligned with the data, it could lead to incorrect agent assignment estimations. This suggestion is further enhanced by the MDR parameters, where even if the aggregated modal split closely matches the one estimated from the reference data, as indicated by the low *mode_distrib_to_work*, the distance distributions are significantly different, as observed in *car_dist* and *pt_dist*. This discrepancy could be due to several factors: inaccurate estimations of the exact route in the initial dataset from the travel survey, which when paired with the Djikstra used for our MATSim simulations, creates misalignments with the initial data; and the aforementioned issue of incorrect OD-matrix assignment in the assignment loop. Nonetheless, not even the calibration through Cadyts in the MATSim simulation could actually fix the distribution of distances performed. Furthermore, the *workers_start_time*, as well as the *worker_act_duration* metrics show a stable trend as the population size increases in both the NRMSE and Hellinger distance, suggesting potential inaccuracies in the initial estimation of these metrics and highlighting one potential drawback of using the OD-matrix sampling assignment.

In conclusion, the analysis demonstrates the effectiveness of the assessment framework in identifying discrepancies between the generated distribution of the parameters of a synthetic population. The results demonstrate that the choice of the percentage of the generated population has a relatively smaller impact on the reliability of the basic sociodemographic and household attributes, while the size of the synthetic population has a clearer impact on activity-travel distance and time distributions. Additionally, the method showed that by increasing the population size, strong discrepancies remain for distance by mode and activity duration, indicating that population synthesizers still have clear room for improvements in matching agents' to activity locations, suggesting that future research should be dedicated to better modeling and constraining activity location choices and trip chaining decisions in order to better match observed daily activity-travel behavior.

CONCLUSION

This paper presented a comprehensive and structured methodology for assessing the performance of the different population synthesis processes aimed at Agent-Based Models (ABM) simulations, with a particular focus on Transportation Oriented Simulations (ABM-TOS). The proposed framework groups different metrics typically analyzed into seven categories, namely: Basic Sociodemographic Attributes (BSA), Household Attributes (HA), Advanced Sociodemographic distributions (ASBD), Trip-chain related distributions (TRPD), Distance-related distributions (DRD), Time-related distributions (TRD), and Mode-related distributions (MRD). The Hellinger distance, a measure sensitive to variations in both the magnitude and the shape of the distributions, is proposed to compare the generated and reference distributions.

The proposed methodology was then applied to a case study involving a novel approach for the generation of a synthetic population for Luxembourg, MOBIUS, using data from a travel survey conducted in the Grand Duchy in 2017. While the results showed that the generated synthetic population closely matched the real-world population in terms of the different categories, it also highlights some of the problems that might be underlying the population synthesis process. The framework proved successful in analyzing and identifying potential problems related to the population synthesis approach while allowing for quick validation and providing an extensive overview of all the parameters that affect the ABMS.

Our findings are crucial for transport planning given the role of a population synthesizer in activity-based models and ultimately city planning and policy appraisal. With the extensive shift in using ABM simulations, without a reliable input, the complicated ABM simulations simply generate misleading output resulting in skewed and spurious policy appraisals. This paper raises concerns and proposes a solution on how the input of AMB should be generated, evaluated and corrected to ensure the ultimate analysis conducted on the ABM-TOS is reliable. In other words, we justified here how dangerous a bad population synthesizer could be and we requested the attention of the community of transport modeling to invest more time and energy in developing reliable population synthesizers.

In future works, we aim to further enhance the precision of our categorization by providing more rigorous rulings for it. This could involve refining the definitions of our existing categories, introducing new categories to capture additional aspects of population synthesis, or developing more sophisticated techniques for assigning synthetic individuals to categories. Moreover, future work could extend the proposed methodology to other case studies and techniques, in order to evaluate the goodness of each one given for a given specific topic. The proposed methodology provides a solid foundation for these future research directions.

ACKNOWLEDGEMENTS

The authors would like to thank the Fonds Nationale de Recherche (FNR), LISER, and the Luxembourg Ministry of Mobility and Public Works (MMTP) for having shared the Luxmobil dataset through the M-EVRST project (Grant Number: 14703944). The authors are also grateful to Dr. Amarin Siripanich for the useful comments and suggestions.

REFERENCES

1. Hörl, S. and M. Balac, Synthetic population and travel demand for Paris and Île-de-France based on open and publicly available data. *Transportation Research Part C: Emerging Technologies*, Vol. 130, 2021, p. 103291.
2. Anderson, A. and A. van Der Merwe, Time-driven activity-based costing related to digital twinning in additive manufacturing. *South African Journal of Industrial Engineering*, Vol. 32, 2021, pp. 37 – 43.
3. Coelho, F., S. Relvas, and A. Barbosa-Póvoa, Simulation-based decision support tool for in-house logistics: the basis for a digital twin. *Computers Industrial Engineering*, Vol. 153, 2021, p. 107094.
4. ODonoghue, C., K. Morrissey, and J. Lennon, Spatial Microsimulation Modelling: a Review of Applications and Methodological Choices. *International Journal of Microsimulation*, Vol. 7, 2013, pp. 26–75.
5. Müller, K. and K. Axhausen, Population synthesis for microsimulation: State of the art, 2010.
6. Farooq, B., M. Bierlaire, R. Hurtubia, and G. Flötteröd, Simulation based population synthesis. *Transportation Research Part B: Methodological*, Vol. 58, 2013, p. 243 – 263, cited by: 101; All Open Access, Green Open Access.
7. Tanton, R., A Review of Spatial Microsimulation Methods. *International Journal of Microsimulation*, Vol. 7, 2013, pp. 4–25.
8. Garrido, S., S. S. Borysov, F. C. Pereira, and J. Rich, Prediction of rare feature combinations in population synthesis: Application of deep generative modelling. *Transportation Research Part C: Emerging Technologies*, Vol. 120, 2020, cited by: 8; All Open Access, Green Open Access.
9. Deming, W. E. and F. F. Stephan, On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals are Known. *The Annals of Mathematical Statistics*, Vol. 11, No. 4, 1940, pp. 427 – 444.
10. Arentze, T., H. Timmermans, and F. Hofman, Creating Synthetic Household Populations: Problems and Approach. *Transportation Research Record*, Vol. 2014, No. 1, 2007, pp. 85–91.
11. Durán-Heras, A., I. García-Gutiérrez, and G. Castilla-Alcalá, Comparison of Iterative Proportional Fitting and Simulated Annealing as synthetic population generation techniques: Importance of the rounding method. *Computers, Environment and Urban Systems*, Vol. 68, 2018, p. 78 – 88, cited by: 6.
12. Beckman, R. J., K. A. Baggerly, and M. D. McKay, Creating synthetic baseline populations. *Transportation Research Part A: Policy and Practice*, Vol. 30, No. 6, 1996, pp. 415–429.
13. Agriesti, S., C. Roncoli, and B.-H. Nahmias-Biran, Assignment of a Synthetic Population for Activity-based Modelling employing Publicly Available Data, 2021.
14. Templ, M., B. Meindl, A. Kowarik, and O. Dupriez, Simulation of Synthetic Complex Data: The R Package simPop. *Journal of Statistical Software*, Vol. 79, No. 10, 2017, p. 1–38.
15. Jain, S., N. Ronald, and S. Winter, Creating a Synthetic Population: A Comparison of Tools, 2015.

16. Tozluoğlu, , S. Dhamal, Y. Liao, S. Yeh, F. Sprei, D. Dubhashi, M. Marathe, and C. Barrett, *Synthetic Sweden Mobility (SySMo) Model Documentation*, 2022.

17. Geyer, C., *Introduction to Markov Chain Monte Carlo*, CRC Press, pp. 3–48, 2011.

18. Felbermair, S., F. Lammer, E. Trausinger-Binder, and C. Hebenstreit, Generating synthetic population with activity chains as agent-based model input using statistical raster census data. *Procedia Computer Science*, Vol. 170, 2020, pp. 273–280, the 11th International Conference on Ambient Systems, Networks and Technologies (ANT) / The 3rd International Conference on Emerging Data and Industry 4.0 (EDI40) / Affiliated Workshops.

19. Sun, L. and A. Erath, A Bayesian network approach for population synthesis. *Transportation Research Part C Emerging Technologies*, Vol. 61, 2015, pp. 49–62.

20. Arkangil, E., M. Yildirimoglu, J. Kim, and C. Prato, A deep learning framework to generate realistic population and mobility data, 2022.

21. Berke, A., R. Doorley, K. Larson, and E. Moro, Generating synthetic mobility data for a realistic population with RNNs to improve utility and privacy. In *Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing*, ACM, 2022.

22. Badu-Marfo, G., B. Farooq, and Z. Paterson, Composite Travel Generative Adversarial Networks for Tabular and Sequential Population Synthesis, 2020.

23. Horni, A., K. Nagel, and K. Axhausen, *The Multi-Agent Transport Simulation MATSim*, 2016.

24. Hanson, S. and P. Hanson, The Travel-Activity Patterns of Urban Residents: Dimensions and Relationships to Sociodemographic Characteristics. *Economic Geography*, Vol. 57, No. 4, 1981, pp. 332–347.

25. Hanson, S., THE DETERMINANTS OF DAILY TRAVEL-ACTIVITY PATTERNS: RELATIVE LOCATION AND SOCIODEMOGRAPHIC FACTORS. *Urban Geography*, Vol. 3, No. 3, 1982, pp. 179–202.

26. Bradley, M. and P. Vovsha, A Model for Joint Choice of Daily Activity Pattern Types of Household Members. *Transportation*, Vol. 32, 2005, pp. 545–571.

27. Hu, Y., B. van Wee, and D. Ettema, Intra-household decisions and the impact of the built environment on activity-travel behavior: A review of the literature. *Journal of Transport Geography*, Vol. 106, 2023, p. 103485.

28. Scheffer, A., R. Connors, and F. Viti, Trip chaining impact on within-day mode choice dynamics: Evidences from a multi-day travel survey. *Transportation Research Procedia*, Vol. 52, 2021, pp. 684–691, 23rd EURO Working Group on Transportation Meeting, EWGT 2020, 16-18 September 2020, Paphos, Cyprus.

29. Chen, Y., *Adding a comprehensive Calibration Methodology to an Agent-based Transportation Simulation*. Ph.D. thesis, 2012.