



**SYNTHETIC GENERATION OF
ACTIVITY-RELATED DATA**

MASTER THESIS

Student

Quentin BOCHUD

Professor

Michel BIERLAIRE

Supervisors

Marija KUKIC
Janody POUGALA

June 2023

Abstract

Synthetic data is increasingly present in our everyday life, like in virtual reality simulations, computer-generated imagery, and algorithmic training datasets. One of their main advantages is their ability to address the issue of the bias present in real-world data and to enhance data quality, especially when dealing with sparse datasets. The field of transportation exhibits a particular interest in these novel techniques due to its strong dependence on expensive survey-based data collection methods.

In this study, we started by selecting the attributes of significant importance in transport habits. Then, we generated list of activity preferences of individuals during one day, using a modified version of Gibbs sampler, a Markov chain Monte Carlo algorithm (MCMC). Our results showed that using only information about the day of the week and individuals' employment status, we were able to generate a sequence of preferred activities that replicated the specificities of a population.

Our work makes a significant contribution by generating preferred activities, even in the absence of this information in the actual sample. Typically, synthetic generation algorithms replicate existing data from a real dataset. However, our approach involves analyzing real activity data to establish a set of assumptions that enable us to generate preferred activities that are not present in the original dataset. These generated activities can then be used as input for existing scheduling models, thereby improving their capabilities.

Contents

1	Introduction	1
2	Literature review	4
2.1	Synthetic activity-related features generation	4
2.2	Daily human mobility motifs	5
3	Theoretical background	6
3.1	Synthetic data generation	6
3.2	Gibbs sampler	6
4	Methodology	8
4.1	Preliminary study and features selection	8
4.1.1	Motif generation	9
4.1.2	Timings study and definition of time slots	9
4.2	Preferred activity generation	10
4.3	Validation	11
5	Results	12
5.1	Data description	12
5.2	Preprocessing	13
5.3	Preliminary analysis and features selection	14
5.3.1	Motif generation	14
5.3.2	Timings and duration	17
5.3.3	Important variables (days and employment)	19
5.4	Definition of time slots	19
5.5	Preferred activity generation	23
5.5.1	Probabilities of preferred activities	25
6	Discussion	29
6.1	Feature selection	29
6.2	Definition of time slots	30
6.3	Preferred activity generation	30
7	Summary and conclusion	31
	References	35
A	Results	36
B	Time slots definition and parameters	38

1 Introduction

In transportation science, transport companies and national mobility offices need to plan the evolution of the mobility of the people. To support these decision makers in creating effective and targeted transportation policies, studies aim to examine the relationship between socio-demographic characteristics and travel behavior by modeling them. By understanding how these factors influence mobility, we can produce more behaviorally realistic models, which can help to guide current and future policy decisions.

In contrast to the 4-step model, which is based on aggregate data and focuses on the four key steps of trip generation, trip distribution, mode choice, and route assignment (Bayes, 2012), activity-based models (ABMs) are used to analyze and predict the travel patterns and behavior of individuals and households (Chu et al., 2012). ABMs are a simulation tool that simulates how people behave when traveling based on their daily activities, including their schedules, destinations, and modes of transportation (Schneider et al., 2013). For this reason, ABMs provide a more detailed and comprehensive understanding of travel behavior by considering individual-level factors and interactions between activities, making them suitable for studying complex travel decisions and policy scenarios. It can be used to evaluate the impacts of different transportation policies or infrastructure investments, and to optimize the allocation of transportation resources. They can also be used to understand the trade-offs that individuals make between different modes of transportation, such as driving, biking, or taking public transportation. In the work of Axhausen (2000), several features related to the travel behavior of the individuals are highlighted such as the kind, the duration, the purpose or the meaning of the activity. These features are useful for understanding the behavior of individuals and are therefore used in multiple models (Felbermair et al. (2020), Hörl and Balac (2021), Pendyala et al. (2012), Pougala et al. (2022)).

To accurately calibrate these models, it is essential to have disaggregated and comprehensive population data that contain information on socio-demographic characteristics and travel behavior. The most common way to collect these data is to conduct surveys, such as population or full activity and travel surveys over one or multiple days for different members of a household or social circle. This is expensive, time-consuming, and can be prone to error (Andrade, 2020). The errors can be from sampling (non-representative sample), non-response (groups of participants refusing to participate, introducing potential bias), measurement (inaccuracies in data measurement or recording) or coverage (excluding certain population segments, leading to underestimation or overestimation). With the arrival of intelligent transportation systems and automated data collection tools (e.g. sensors), the amount of available data has drastically increased. However, current regulations limit the usage of the data that includes sensitive and private information (Stopczynski et al., 2014).

For these multiple reasons, research is increasingly turning to the usage of synthetic data rather than real-world collected data. In general, one of the main advantages of utilizing synthetic data stems from the ability to generate it by combining various data sources. This approach can be beneficial in addressing the issue of bias present in real-world data or enhancing data quality, particularly when dealing with sparse datasets. The privacy protection and costs saving are other advantages. In the literature, different algorithms for synthetic data generation can be found and are grouped in three main categories (Yaméogo et al., 2021). These categories are synthetic reconstruction (SR) (Beckman et al., 1996), combinatorial optimization (CO) (Ma & Srinivasan, 2015) and statistical learning (SL)

(Sun et al., 2018). SR and CO methods produce synthetic populations by means of replicating individuals, whereas SL methods generate a population following a joint probability estimation. The choice of the algorithm depends on the user needs and available resources. One of the key challenges in generating synthetic data is ensuring that the data is realistic and representative of real-world data, to ensure that models trained on this data are relevant and reliable.

One of the purposes of transport data is to analyze schedules of a population. As described in Damm and Lerman (1981), an activity-schedule is the plan of the different activities of an individual's day with the start times and durations of these activities. In Pougala et al. (2022), real-data are used and a new approach to modeling daily activity scheduling is made. This approach considers the multiple dimensions of the schedule as a single optimization problem. The objective is to generate schedules that maximize the global utility derived from an individual's activities, considering his or her needs, constraints and preferences. To recreate schedules, Pougala et al. use activity participation, timing, activity sequencing, location and mode of transportation collected via a micro-census. However, the work presented in this paper is limited by the use of an actual travel survey, which lacks important information such as desired schedules, unchosen alternatives, or schedules of other household members. To overcome these limitations, a synthetic data can be used which would consequentially extend the capabilities of the method of Pougala et al. Therefore, the goal of this master project is to provide data that are not available in the real dataset.

In order to generate these data, we adapt the methodology introduced by Kukic and Bierlaire (2022). This methodology uses Markov chain Monte Carlo (MCMC) algorithm to generate synthetic households. They use so-called one-step Gibbs Sampler that iteratively draws from conditional distributions provided as input to approximate a multivariate distribution formed by desired attributes. This methodology simulates a hierarchical structure, such as individuals grouped into households. Compared to the original methodology, we change this hierarchical structure to individuals with set of preferred activities they would like to perform. To describe these sets, we generate a sequence described by activity type, start time and activity duration. In case of household generation, the order of the individuals is not important, while in the generation of sequences of preferred activities and start times, we have to make sure that they are generated in a specific order. Moreover, the duration of activities differs among different individuals, so we have to make sure that we generate a flexible length of activities sequence. Since the original methodology deals with discrete variables, in our work we expand this method to include continuous variables such as duration or start time.

This master project was preceded with a pre-study to define its goals and scope. A literature review was conducted to understand the subject and the current state of the field. The Gibbs sampler algorithm was studied based on the work of Kukic and Bierlaire (2022). Following that, various applications of the Gibbs sampler were tested for both discrete and continuous attributes. Generating discrete attributes proved successful, closely replicating the original values. However, generating continuous attributes posed challenges, resulting in deviations from the original distribution. The pre-study highlighted the need to develop a method for generating time values that closely resemble the original data and to manage data having a link (schedules for example). Data validation and handling interconnected relationships were identified as important considerations and rigorous methods for data validation were recommended for the master project.

Following the context and the motivations explained above, the aim of this master project is to address two research questions :

- How to represent and capture a population's most popular activities, including start times and duration ?
- Using Gibbs sampler, how to generate synthetically the preferred activities ?

For this purpose, the current report is composed as follows. Following a review of the literature in Chapter 2 and a theoretical background (Chapter 3), a definition of the methodology is made (Chapter 4). Then, the case study is presented as well as the results (Chapter 5). Finally, a discussion of the results in Chapter 6 precedes the conclusion in Chapter 7.

2 Literature review

This literature review examines two key subsections: the generation of synthetic activity-related features and the study of everyday human mobility motifs, which consists of sequences of activities performed or locations visited by individuals. We explore recent advances in these areas, highlighting methods, results, limitations and opportunities.

2.1 Synthetic activity-related features generation

Understanding people’s preferences for certain activities at different times of the day is essential for more accurate and responsive transportation planning (Rich et al., 2021). As explained above, the use of synthetic data is becoming essential. The field of synthetic data generation is vast, but the most central is population generation, which can be used as a basis for generating other synthetic data. Traditional models for generating synthetic population such as Beckman et al. (1996), Arentze et al. (2007) or Pendyala et al. (2012) typically rely on travel surveys and statistical models to generate a synthetic population. The use of demographic characteristics such as age, gender, income level, and household composition to model individuals is common (Müller & Axhausen, 2010).

However, these approaches are often limited by their inability to capture individual variations and complex interactions between different activities (Bae et al., 2020). In addition, they may lack the flexibility to represent changes in preferences over time and in specific contexts. For example, traditional models may not take into account the fact that individuals are more likely to go shopping after leaving work or to engage in leisure activities at weekends. Furthermore, these models are not always capable of representing temporal variations in mobility behavior, particularly with regard to trips made at different times of the day (Hörl & Balac, 2021).

To overcome these limitations, machine learning techniques such as in Felbermair et al. (2020) have been used to generate synthetic activities. These methods exploit sophisticated algorithms such as neural networks or hidden Markov processes to model the complex relationships between explanatory variables and travel or activity choices. In Farooq et al. (2013) for example, a population synthesis methodology based on Markov chain Monte Carlo (MCMC) simulation is used. These models have shown significant improvements in terms of dependency representation. However, the use of machine learning-based methods often requires large, high-quality datasets for training (Borysov et al., 2019). The use of real mobility data from surveys or sources such as transit meters can help improve the quality of the synthetic data generated. In Felbermair et al. (2020), the use of Bayesian networks and MCMC as well as stratified sampling, showed how a population with activity plans can be generated using limited survey data. In Kitamura et al. (1997), they propose a micro-simulation approach to generate synthetic daily activity-travel patterns with a sequential modeling that decomposes the entire daily activity-travel pattern. Here, they generate defined activities, but the probability of doing an activity at a defined time of the day is missing. As explained in the introduction, this feature is required for scheduling studies such as Pougala et al. (2022) and the current master project aim to address this lack.

To gather data, the use of location-based data and technologies such as mobile devices and social networks opens up new opportunities for generating preferred activities according to the time of day. These data sources enable to capture the actual behaviors of individuals in their daily environment, offering richer and more accurate information for activity modeling. However, it must be borne in mind that this information raises the issue of privacy. For example, in Berke et al. (2022), they used a system for generating synthetic mobility data using a deep recurrent neural network (RNN) to solve this problem.

Evaluating and validating models for generating synthetic transportation and preferred activity

data can be complex. It can be difficult to determine the extent to which the data generated corresponds to actual observed behavior (Ma & Srinivasan, 2015). Rigorous methodological approaches and thorough validation studies are needed to assess model performance and ensure their relevance. The main challenge of synthetic data generation is to succeed in developing models that take into account individual preferences, temporal constraints and interactions between activities, and therefore better capture the complexity of real behavior. The use of MCMC methods, such as Gibbs sampler discussed below, can help overcome these limitations.

2.2 Daily human mobility motifs

Although the human behavior is diverse, some studies such as Su et al. (2020) or Cao et al. (2019) show that almost all human movements can be aggregated and reduced to several location or activity based motifs. In Schneider et al. (2013), a motif is defined as a directed graph, in which nodes represent visited locations or practiced activities and directed edges (i.e., links) represent trips between locations or activities. For this reason, this is an interesting tool for studying the mobility habits of individuals, as it enables behaviors to be compared (Schultheiss et al., 2021). All of these papers show that individuals follow consistent patterns of activity according to the group to which they belong, and that it is possible to analyze mobility behavior under the spectrum of motifs. For example, Ectors et al. (2019) show that very few activities are very frequent, while others are rare. They additionally prove that activity-based motifs follow so-called Zipf’s law, where the most frequent motif is about twice as frequent as the second, three times as frequent as the third, and so on.

Activity or location based motifs can be valuable for synthetic generation for two reasons. Firstly, the motifs do not reveal the exact location of the individual which do not violate the privacy constraints. Secondly, we can use motifs to aggregate data before generating them. This reduces the complexity, instead of trying to reproduce all patterns that exist in reality, we can generate the significant motifs. This can contribute to achieving the same representativity of the synthetic movements in more efficient manner (Ectors et al., 2022).

3 Theoretical background

In this chapter, we first describe synthetic data algorithms to generate populations and then, the Gibbs sampler algorithm is presented.

3.1 Synthetic data generation

There are several existing methods for generating synthetic data, including Iterative Proportional Fitting (IPF), Markov chain Monte Carlo (MCMC) techniques, and machine learning methods such as generative adversarial networks (GANs) and variational autoencoders (VAEs).

IPF is a method for generating synthetic data that is based on the idea of iteratively adjusting the margins of a contingency table to match known marginal totals (Beckman et al., 1996). This method can be used to synthesize data for a wide range of variables, including both continuous and discrete variables. IPF is a popular method for generating synthetic data because it is relatively simple to implement and can produce synthetic data that is consistent with real-world data (Bierlaire et al., 2021).

MCMC techniques are a class of algorithms that are used to estimate the parameters of a statistical model by sampling from the model's posterior distribution. These techniques can be used to generate synthetic data by sampling from a probability distribution that is defined by the statistical model. MCMC techniques are widely used in statistical inference and are popular for generating synthetic data because they can produce accurate and reliable estimates of the model parameters (Bierlaire et al., 2021).

GANs and VAEs are both types of machine learning models that can be used to generate synthetic data. GANs consist of two neural networks, a generator, and a discriminator, that are trained to compete against each other. The generator tries to create synthetic data that is similar to a training dataset, while the discriminator tries to distinguish the synthetic data from the real data. Through this process, the generator learns to generate synthetic data that is similar to the real data (Vieira, 2020). VAEs are a type of generative model that consists of an encoder and a decoder. The encoder takes in the real data and encodes it into a latent representation, and the decoder takes this latent representation and generates synthetic data that is similar to the real data. Both GANs and VAEs can be used to generate synthetic data that captures the structure and patterns in the training data and can be useful for tasks such as data augmentation and creating simulated environments. However, they can be challenging to train and may require a large amount of training data to generate high-quality synthetic data.

3.2 Gibbs sampler

The Gibbs sampler is a MCMC algorithm that is used to approximate a multivariate distribution. It works by iterative sampling from the conditional distributions of each variable, given the current values of the other variables. This process is repeated for a sufficient number of iterations, after which the samples can be used to approximate the joint distribution of the variables. In Bierlaire et al. (2021) the process is explained as follows:

1. **Initialization** Choose initial values for all of the variables.
2. **Warm up** Generate a sequence of individuals and do not include them in the population. The procedure to draw a value is as follows:

Suppose we have a set of variables x_1, x_2, \dots, x_n and we want to sample from their joint distribution $p(x_1, x_2, \dots, x_n)$. The Gibbs sampler works by iteratively sampling from the full conditional distributions of each variable, given the current values of the other variables.

For each iteration of the algorithm, we do the following :

- (a) Select randomly a variable x_i to update.
 - (b) Given the current values of the other variables $x_1, x_2, \dots, x_{i-1}, x_{i+1}, \dots, x_n$, sample a new value for x_i from its full conditional distribution $p(x_i \mid x_1, x_2, \dots, x_{i-1}, x_{i+1}, \dots, x_n)$.
 - (c) Update the value of x_i with the new sample.
3. **Populate** When the warm-up phase is finished, we generate the next individual from the sequence and include it in the population with the same method as in the warm-up phase.
 4. **Skip** Generate a sequence of individuals using the procedure described above, and do not include them in the population.
 5. **Iterate** Repeat the steps “Populate” and “Skip” until the generated population contains enough individuals.

The Gibbs sampler algorithm offers great flexibility in modeling the complex probability distributions associated with preferred activities as a function of the time of day. It allows the parameters of the model to be adjusted according to the input data, resulting in more accurate results tailored to the specific characteristics of the population under study (G.O. Roberts, 1994). The main contribution of the Gibbs sampler is that we do not need disaggregated data as an input. We can use only marginals to generate what we want, which implies that we can generate data without having a sample.

4 Methodology

The goal of this master project is to generate probabilities of what the people would like to perform at a certain time. For this purpose, Gibbs sampler is used. The standard Gibbs sampler for activity generation generates one sequence of activity types that are performed at the specific time of the day with a specific duration (Kukic & Bierlaire, 2023). This methodology can be modified to generate a sequence of possibilities of what people would most likely do, based on the observed data for each part of the day. We call this sequence of probabilities per time slot a list of preferred activities. The general methodology of this project is shown in Figure 1.

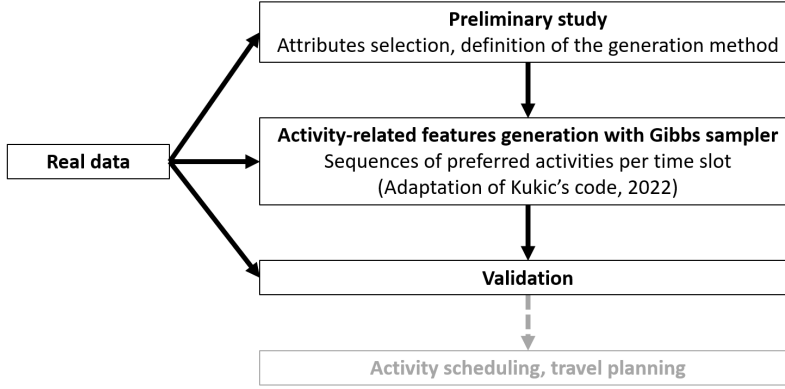


Figure 1: General methodology

The first step of the general methodology is to conduct a preliminary study to select the minimal set of attributes from the real-data set that are mandatory to be included in the generation process and that mostly influence the person's schedule. We divide the generation process into different subsets depending on the values of selected features. This allows us to see which attributes have an impact on the activities performed by the individuals (e.g., distribution of start time or motif). Once these influent attributes have been identified, the methodology for the preferred activities generation can be defined. After the generation, we propose how to validate the generated preferences before the data can be used by other studies, such as activity scheduling for example.

4.1 Preliminary study and features selection

To see the determinant attributes, we want to select socio-demographic or timing attributes (e.g. employment status or day of the week) that significantly impact the activities performed by the individuals. In the following Figure 2, the detailed process is presented.

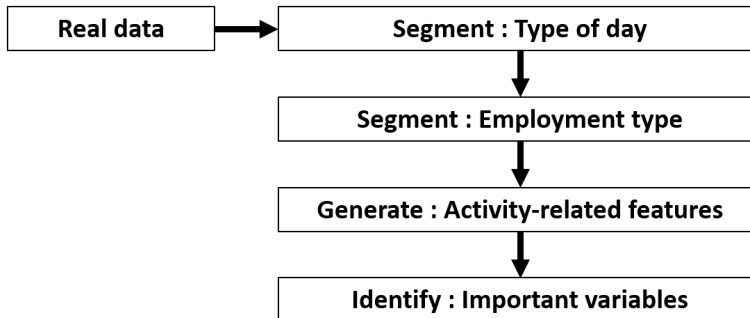














Figure 2: Data segmentation

From a real dataset, we separated the data according to the day type (weekdays and weekend days) and then, each of the two groups is separated according to employment status (full-time worker, part-time worker, student, or unemployed). We do this segmentation because we assume that the type of activities performed will be different depending on the type of day and that the employment status influences the activity behavior of the individuals. Following this sampling, activity-related features are generated and their distribution are analyzed. These features are start time, durations and sequences of activity type (motifs). This first analysis allows us to check the general coherence of the data by confirming the importance of these variables (day of the week and employment status).

4.1.1 Motif generation

To identify what are the potential preferences of an individual, we investigate their habits. To do this, we can design activity-based (Schultheiss et al., 2021) or location-based motifs (Cao et al., 2019). Instead of observing the disaggregated data about the trip diary of one individual, all the activities the person performed during one day are aggregated into motifs. In the original methodology of Schultheiss et al., we can choose to extract either location-based or activity-based motifs. In each node, we also store the information of the start time and the duration of each activity. Here, we create activity-location motifs. This implies that if the two different activities are performed at the same locations, we create only one node. Moreover, in Schultheiss et al., if two motifs have the same shape, they are considered to be the same (e.g, home-work-home and home-leisure-home), while in our work, they are treated as two different motifs. Table 1 describes the differences in the generation between the three methods.

Table 1: Differences between motifs definitions

Original activities	Activity-based model	Location-based model	Activity-location based model
			
			
			

Once the motifs are extracted, we analyze their distribution across the whole sample. Since we noticed that the most of the motifs are specific for a group of the population, we investigate further which socio-demographic features impact the shape of the motifs. This allows us to capture the characteristics of a population group.

4.1.2 Timings study and definition of time slots

Having analyzed the population’s motifs, it is now essential to be able to capture these activities while retaining information on the start time and duration. The solution is to divide the day into parts and then assign an activity to each “slot”, as in Kukic and Bierlaire (2023). Based on their work, the method is modified to define time slots in several different ways. To decide how to define time slots,

we analyze the typical start times and durations for activities across the categories of the population. To do this, from the motif of each individual, we extract the start time and duration of each activity and analyze their distribution for different population groups.

After having analyzed the characteristics of start times and durations, time slots can be defined. The three final methods are described below :

1. Method 1 is to define time slots arbitrarily according to expert knowledge dictated by the results of the analysis on the timings. This method is based on the one originally used in Kukic and Bierlaire (2023). In their work, the time slots are defined the same independently of the type of individual or the day of the week. Here, a modulation according to the importance of the socioeconomic attributes and the day of the week allows to be more realistic.
2. Method 2 is to automate time slot definition by isolating the start time data peaks. Here, the method is based on the groupings of start times. It is an automatic and more precise extension of the first method. Time slots are defined in the same way, depending on the socioeconomic attributes and day of the week, but automatically.
3. Method 3 consists of defining a number of time slots with a constant duration. This method defines time slots in the same way, whatever the day of the week or the type of individual. The number of time slots and their duration can be modulated.

For all these methods, the procedure remains the same each time. Once the time slots have been defined, each activity is assigned the time slot(s) during which it takes place. Several activities of the same individual can therefore have the same time slot assigned. An example of slot assignation can be seen in Figure 2. After this, the validation method (see Section 4.3) for assessing the quality of time slot definition is used to see which method is most effective at capturing preferred activities. These time slots are then used for generation using the Gibbs sampler.

Table 2: Example of time slots assignation

Starting time	Activity	Time slots		
		Method 1 weekdays Full-employed	Method 2 weekdays Full employed	Method 3 24 slots of 1 hour
00:00	Home	1, 2	1, 2	0, 1, 2, 3, 4, 5, 6, 7
07:15	Work	2, 3	2, 3	7, 8, 9, 10, 11
11:57	Home	3	3	11, 12, 13
13:21	Work	3, 4, 5	3, 4	13, 14, 15, 16, 17
17:32	Shopping	5, 6	4	17, 18, 19
19:03	Home	6	4	19, 20, 21, 22, 23

4.2 Preferred activity generation

To generate preferred activities per time slot for each individual type, we apply a modified version of the Gibbs sampler algorithm used in Kukic and Bierlaire (2022). The general methodology is explained in Section 3.2 and as discussed earlier, the aim of this work is to modify it to generate preferred activities per time slot. With the standard method, the output of Gibbs sampler is the activities performed (e.g., [Home, Work, Home]) and with our approach, the probabilities of the activities are generated (e.g., [[Home 100 %], [Work 80 %, Home 10 %, Leisure 10 %], [Home 100 %]]).

To generate these probabilities, three attributes are selected: employment status, time slot and activity. Employment status is also used in Kukic and Bierlaire, and time slots and activity are the

attributes we want to generate in the end. Conditional tables are formed by having one attribute conditional on the other two. We assume that there exist differences between weekdays and weekends, therefore, we model these separately. Then, during the warm-up and the following phases, each attribute is randomly drawn and its new value is defined from its conditional distribution. The final methodology for generating preferred activities is summarized in Figure 3.

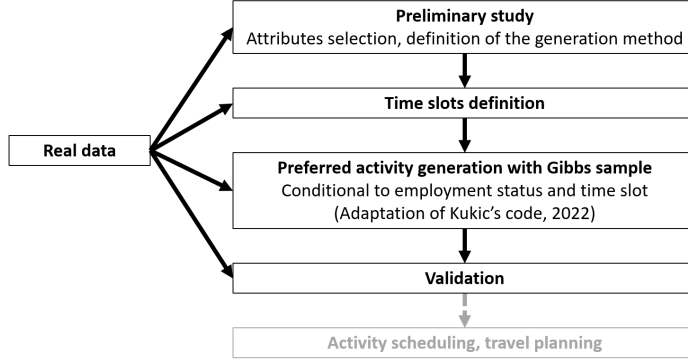


Figure 3: Methodology for preferred activities generation

4.3 Validation

The most common validation technique for synthetic data is to compare different distributions against the real ones by visualization or statistical tests. If the distributions are similar, we consider the synthetic data representative. To test the quality of the different definitions of time-slots, we extract the most popular activity type per time slot and compare these activities with the activity-based motifs.

This method involves examining the most popular activities by time slot and comparing these activities with the activity-location based motifs generated previously. By doing this, it is possible to capture a sequence of the most popular activities during the day for a given population type. In mathematical terms, it means that the sequence of all the activities with the higher frequency in each time slot is obtained. This sequence can be compared with the most frequent motif in the original data. Depending on how the time slots are defined, it is possible to capture activity sequences more or less well, and consequently to validate or invalidate certain definition methods.

5 Results

In this section, first, the data description and the preprocessing are presented. After this, the scenarios used to judge the quality of the model and the best way to generate data are shown. Finally, the results of the preliminary study, evaluation of the time slot definitions, and the activity generations are given. For the sake of clarity, we only discuss selected results in this section. All the codes and results are available on the GitHub repository.

5.1 Data description

The data used in this report are from the Swiss Mobility and Transport Micro Census Data 2015 (MTMC), collected by Federal Office for Spatial Development (ARE) and the Federal Statistical Office (FSO). This data sample gathers information on people's mobility behaviors. More precisely, respondents list their socioeconomic features, their daily mobility routines (such as time or distance to work), and detailed records of their travels throughout a reference period (1 day). Originally, a total of 279'174 trips are represented in the disaggregated sample. This dataset is chosen to be aligned with the results of studies by Kukic and Bierlaire (2021) and Pougala et al. (2022).

The description of the data used in this case study is given in Table 3. The chosen attributes describe the socio-demographic of individuals (e.g., age, gender) and activity attributes (e.g., start time, end time, duration, motif type).

Table 3: Data description (*: continuous attributes)

	Name in the dataset	Meaning
From MRMT 2015	f51100	Start time of the trip*
	f51400	End time of the trip*
	f52900	Purpose of the trip (activity)
	HHNR	Household number
	ERWERB	Employment status (full-time, student, etc...)
	f41100_01	Professional position (manager, employee)
	f81000	Working domain (primary and secondary sector)
	f81100	Working domain (tertiary sector)
	tag	Day of the week
	alter	Age of the individual*
	gesl	Gender
	f41610a	Owner of the GA travelcard
	zivil	Civil status (single, married, etc...)
After preprocessing	started_at	Start time of the activity*
	finished_at	End time of the activity*
	duration	Duration of the activity*
	motif_type	Motif generated

Travel and socioeconomic attributes are directly available, while activity attributes are preprocessed as described in Section 5.2. Motifs are generated according to the methodology described in Section 4.1.1. The start and end times of activities are respectively associated with the end time of the previous recorded trip and the start time of the next. Finally, durations are extracted from activity starts and ends.

The activities generated and selected are defined according to MTMC 2015. Table 4 illustrates the

list of all possible activity types.

Table 4: Activities description

Number in the dataset	Abbreviation	Meaning
0	H	Home
1	C	Change, change of means of transport, turn off the car
2	W	Work
3	E	Training, school
4	S	Shopping
5	Se	Contributions and use of services (e.g., hospital)
6	B	Business activity
7	R	Ride
8	L	Leisure activity
9	A	Accompanying path (only children)
10	A	Accompanying path/service path (others, e.g., disabled people)
11	H	Return to your home or outside accommodation
12	O	Other
13	Bc	Border crossing
-99	-	Pseudo stages
-98	-	No Answer
-97	-	I do not know

5.2 Preprocessing

Prior to the feature selection, the data needs to be cleaned. A number of activities or individuals are removed from the dataset following assumptions described in Table 5. Note here that the assumptions and the preprocessing method are the same for the feature selection and the generation process.

Table 5: Assumptions

Assumptions	Data (%)
Do not consider activities at home (set home purpose)	-
Combine all mode changes into a single activity	-
Filter out users that have only one activity	7%
Filter out users that do not start or finish their day at home location	8%
Filter out users that have at least one trip across Swiss borders during the day	7%
Filter out users with overnigh activities	< 1%
Filter out under 6 min activities	6% of the activities
Total activity lines	177'468 lines (64%)
Total individuals	38'971 households (77%)

The assumptions are made in accordance with the study of Schultheiss et al. (2021). Since this study focuses on understanding travel-related motifs, the motifs that involve only staying at home are excluded. We want to have routine motifs, and this is why the people who do not start or finish their day at home are removed. The analysis focuses on journeys within Switzerland, which is why a geographical limitation is applied. Finally, activities over multiple days are not taken into account, because analysis focuses on activities performed during one-day. After the application of

the assumptions, 77% of the original individuals remain (38'971 household members), and 177'468 activity lines (64%) are considered.

5.3 Preliminary analysis and features selection

In the preliminary analysis, we want to see what attributes have the biggest impact on the behavior of the individuals. The generation of activity-location based motifs (sequence of activities) and timings features are performed.

5.3.1 Motif generation

Following the methodology described in Section 4.1.1, we split the original dataset (MTMC 2015) according to the type of day and then, into different employment status categories. Figure 4 shows the 10 most common motifs for the whole sample before distinguishing the day of the week or other socioeconomic attributes.

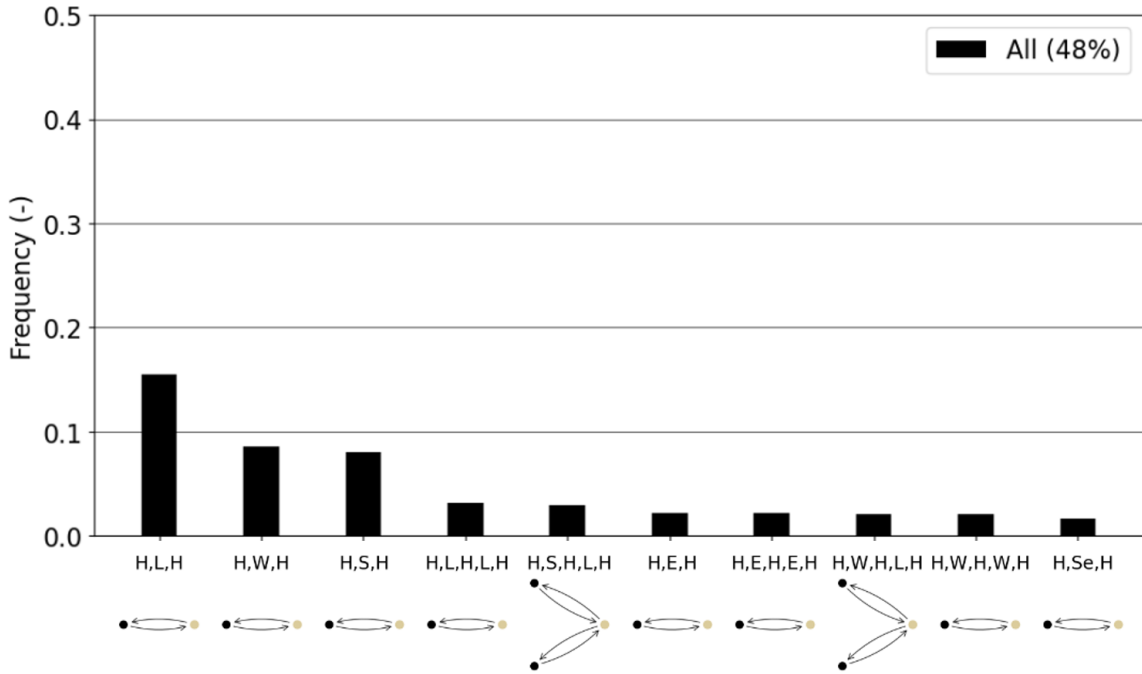


Figure 4: Distribution of the motifs for the whole population

Firstly, we can see that 48% of the out-of-home schedules are represented in the top 10 motifs. As a reminder, we removed 7% of the data of the people who stayed at home all day, because we want to focus the analysis on the out-of-home activities. 48% is less than in the other studies such as in Schneider et al. (2013), where 90% of the data is represented in the first 17 motifs. The main reason for this difference is the fact that we do not group the motifs with the same shape but with different activities, such as in Cao et al. (2019) for example. The variety of motifs is therefore higher. In this top 10, three different types of motifs are generated. First, there is the basic home-xx-home (motifs 1, 2, 3, 6, 10). Then, there is the variant with a return trip home during the day (4, 7, 9), and finally, one with a second activity after returning home (5, 8). We can see that from the fifth motif, the percentage is already below 3% and that the four most common motifs (35% of data) concern leisure, work, or shopping activities. We can see here that a refinement of the analysis is requested by adding parameters to see trends. Therefore, the data is split between weekdays and weekends and

the results can be seen in Figure 5.

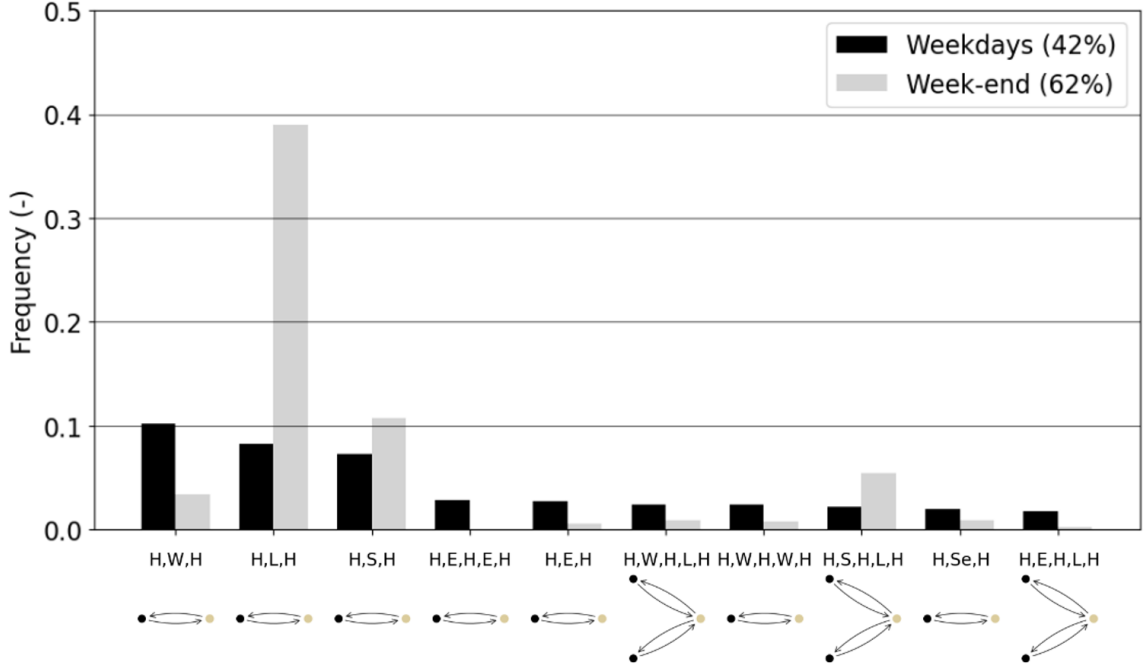


Figure 5: Distribution of the motifs depending on the day of the week

As expected, there is a clear difference between the two types of day. The data represented by the top 10 motifs is 42 % for weekdays and 62 % for weekend days. We note here that activities are less varied at weekends than on weekdays (more data represented in the top 10), and that the home-leisure-home motif is strongly practiced at weekends (39 %). Motifs involving work or education (1, 4, 5, 6, 7, 10) are undoubtedly more present on weekdays than at weekends. It should also be noted that the shapes of motifs are the same as in the previous Figure 4.

After separating the dataset regarding the days of the week, we split it according to the employment status of the person. In Figure 6, the four employment statuses (full-time employee, part-time employee, student, unemployed) with the 10 most common motifs per category are shown.

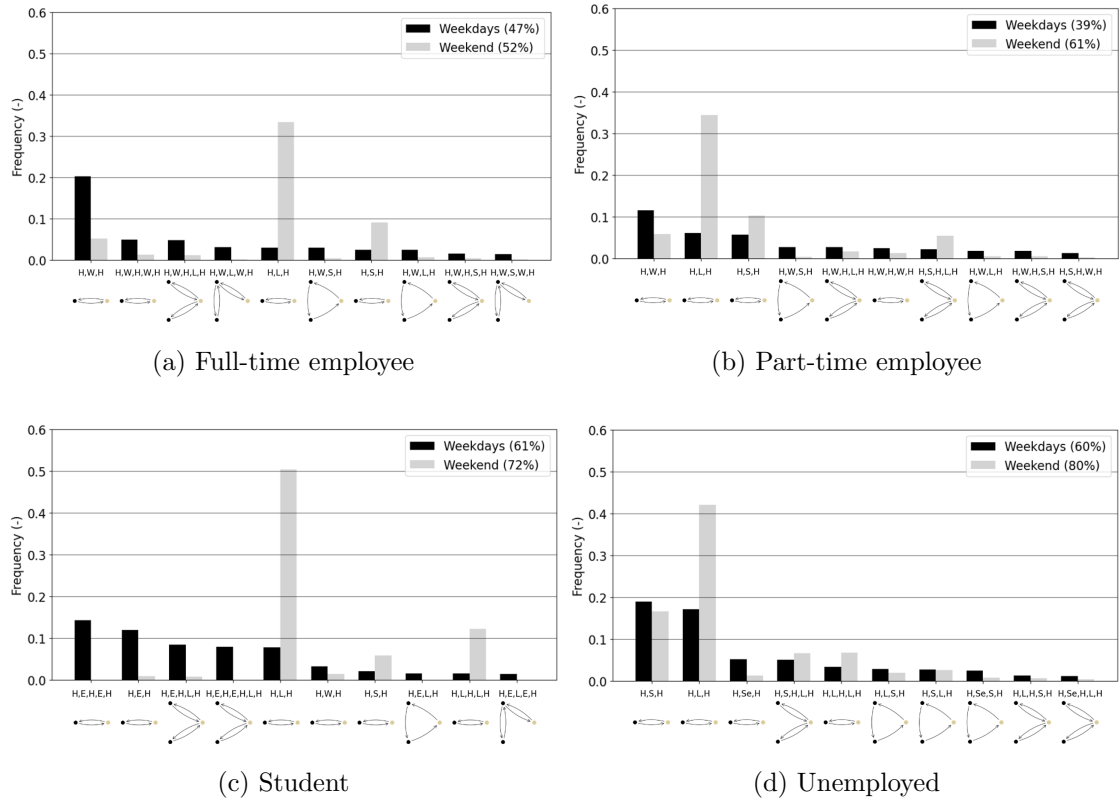


Figure 6: Distribution of the motifs depending on the employment status and the day of the week

First of all, we can see that for each type of population, motifs for activity are more varied during the weekdays than at weekends (percentages of motifs present in the top 10 are systematically lower on weekdays). The activity of leisure is the most performed during weekends (for every employment category has a percentage of the home-leisure-home activity above 30 %).

For people categorized as working (full-time or part-time), the data represented by the 10 most frequent motifs is below 60 %, even on weekends. For full-timers on weekdays, home-work-home is the most frequent motif with 20 %, and it is completed by 5 % of people going home during the work period. For those working part-time, the home-work-home motif accounts only for 12 %, but they have other motifs, notably leisure or shopping only. We can see here the differences between the types of workers, with more varied and less work-only oriented activities for part-timers, than for full-time workers.

People in education are those with the most marked differences between weekdays and weekends. On weekends, motifs involving study-related activities are nearly absent, whereas they represent the majority of weekday motifs. This is due to the way institutions in Switzerland teach (no or few courses at weekends). These people have motifs particularly common to their population at weekends, with three types of motifs accounting for 69 % of all motifs. These motifs involve leisure or shopping activities. This is due to the fact that students are for the most part still young, and do not have the same constraints as other population groups, who do not have the whole weekend to do only leisure activities.

Finally, people categorized as unemployed are also significantly different from other population groups. This category represents people who are at home and have to look after the household. Here, the motifs and probabilities are similar between weekdays and weekends. The only major difference is the higher probability of leisure activities at weekends. This shift in probability is partly due to use of services (Se), which are much less present at weekends than on weekdays. With a top 10 representing over 80 % of all weekend motifs, this category has the least variety of motifs.

To sum up, there are clear differences between the different days of the week and between the different employment statuses. This may seem obvious, but it does highlight the variations between categories (e.g., mainly education activities for students during weekdays or a majority of leisure activities during weekends). We can already see at this stage that it will be important to take these attributes into account when generating activity-related features.

5.3.2 Timings and duration

To investigate further the differences between employment status categories, we observe the distribution of the start time and duration of activities. Another attribute that can help to understand the mobility behavior is the frequency of each activity per day. For example, one activity can be performed multiple times per day and the duration can vary depending on whether it is performed for the first, second, or third time of the day. This information is denoted as “occurrence” in the graphs. For example, if an individual performs the activity “work” for the second time in the day, the occurrence of this activity will be 2. In brackets is written the total number of activities for the day at that occurrence. In Figure 7, we see the start times of the activities “work”, “school”, “shopping” and “leisure”.

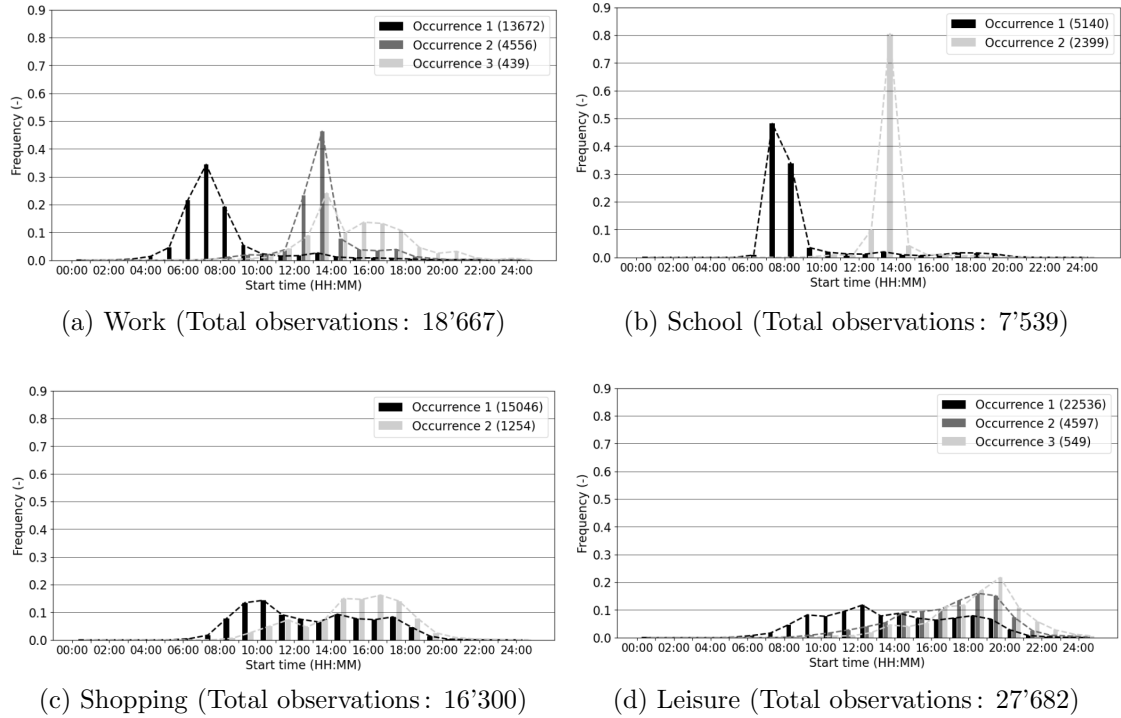


Figure 7: Start times of the activities

In the case of work or school, the peaks are clearly marked, and it is easy to see when people start these activities. On the other hand, for leisure or shopping, there is no specific time, and the distribution is spread out over the whole day. If we look specifically at these activities, we can highlight some ranges when they are performed. For shopping, we can see that purchases are made between 07:00 and 20:00 which corresponds to store opening hours. For leisure, the range extends later in the day, since it is not obligatory linked with opening hours and depends more on the willingness of the people.

For work, the first occurrence of the day has a clear peak between 06:00 and 09:00. This corresponds to people who start to work in the morning. The second occurrence peaks just after the lunch break, between 12:00 and 14:00. For students, the peaks are even more pronounced and do not spread out. The first occurrence is between 08:00 and 09:00 (in Switzerland not all schools are at exactly the same time). The second one is even more marked with 80 % of students starting again between 13:00 and 14:00 for the afternoon classes.

Regarding the number of times a person engages in each activity during the day, we can see that school and shopping activities are carried out no more than twice a single day for the same individual. On the contrary, we can see that some people perform work and leisure activities up to three times.

To be able to decide where the time slot start and end times are, we must look at activities start times, but also at activities durations. It is the combination of the two that determines the time slots of method 1. The duration of the same activities as in Figure 7 can be seen in Figure 8.

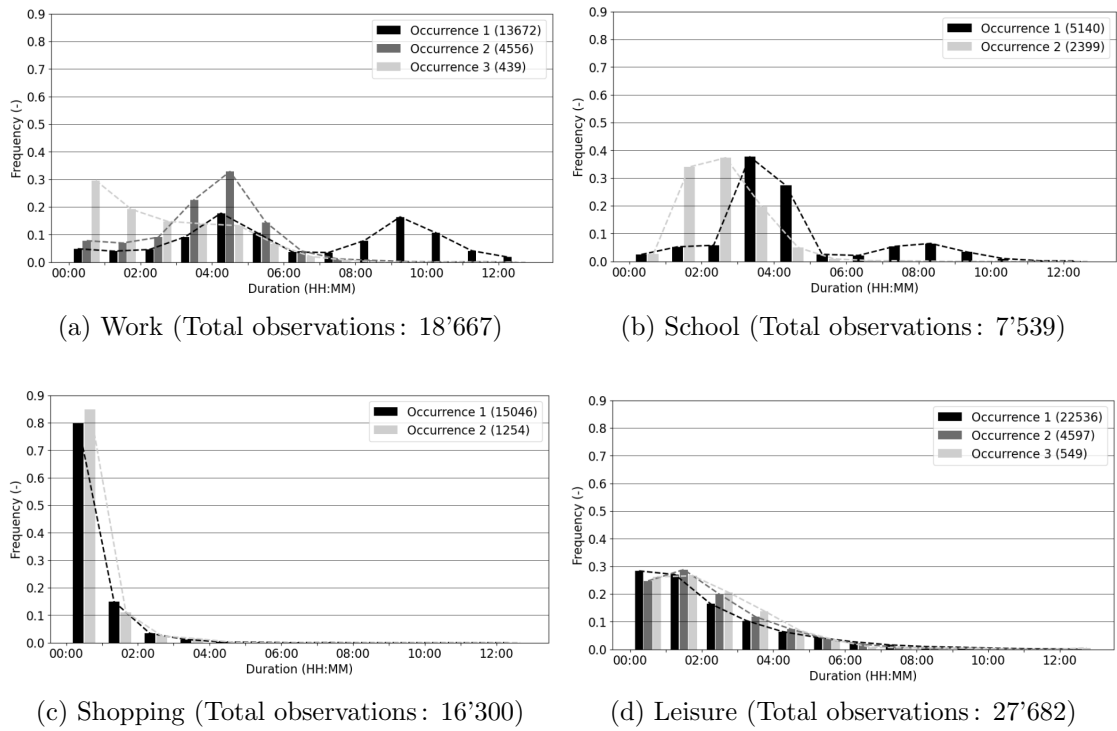


Figure 8: Durations of the activities

The activity duration results are in line with expectations. We can see that there are big differences depending on the type of activity. Generally speaking, work and school activities can extend longer into the day than leisure activities. This is even more striking with shopping activities.

If we look at the work activity, we see that for the first occurrence, there are two distinct peaks (4 and 9 hours). The first corresponds to people who only do half a working day, while people who do the whole working day have a peak duration of around 9 hours. The second occurrence of the day also lasts 4 hours, which corresponds to the half working day. If there is a third occurrence, it has a shorter duration on average than the first two, which corresponds to the after-dinner working slot.

For school activity, we find the same pattern as for work, with two distinct peaks for the first occurrence of the day. The second peak is less pronounced than the first, because as we saw in Figure 7, there are many students who do not spend the whole day studying. We also note that the duration

of the second occurrence is shorter than the first. This refers to the afternoon classes, shorter than the morning ones. Note here that nearly no activities last between 5 and 7 hours because it would finish after the lunch break, if it started at 07:00, which is not common.

For shopping and leisure activities, the number of occurrences does not seem to influence duration. For shopping, activities are short (mostly under an hour), while for leisure, activities have longer durations (up to 7 hours). The duration of shopping activities is not surprising because people mostly stay less than an hour in shopping centers, and if they go to another place, the activity is considered different in our model.

5.3.3 Important variables (days and employment)

Following the preliminary study, we can conclude several things. Firstly, motifs vary according to the day of the week and the employment status of the individual. On weekdays, workers are mostly engaged in work activities, with a high number of different motifs, and it is difficult to pinpoint a motif other than the classic home-work-home (20 %). The activity-location based motifs of students during the week are clearly focused on education. At weekends, the activities of the different categories of people are more similar and mainly oriented toward leisure activities. This should be borne in mind when validating the activities generated.

Secondly, activity start times and durations also show that there are differences between categories of people depending on their employment status and type of activity. The results show that work and school activities had clear, well-marked start times for the entire population engaged in these activities. Activity durations are also in line with what might be expected (e.g., short shopping duration, differentiation in duration between the first and second work or school activities, specific activity duration curve for leisure activities, etc.). These are important points to bear in mind when defining time slots with method 1 and in the activity generation process.

5.4 Definition of time slots

To define the time slots, the three methods described in Section 4.1.2 are used. Concerning method 1, which consists of defining the time slots arbitrarily and according to expert knowledge, the time slots are defined as illustrated in Figure 9. Differences are made according to the day of the week and employment status following the study of these attributes conducted previously. For the weekend days, only the student category has different time slots, because of the specifications highlighted in the preliminary study. For weekdays, full-time and part-time workers have the same time slots definition, because of the similarity of their motifs.

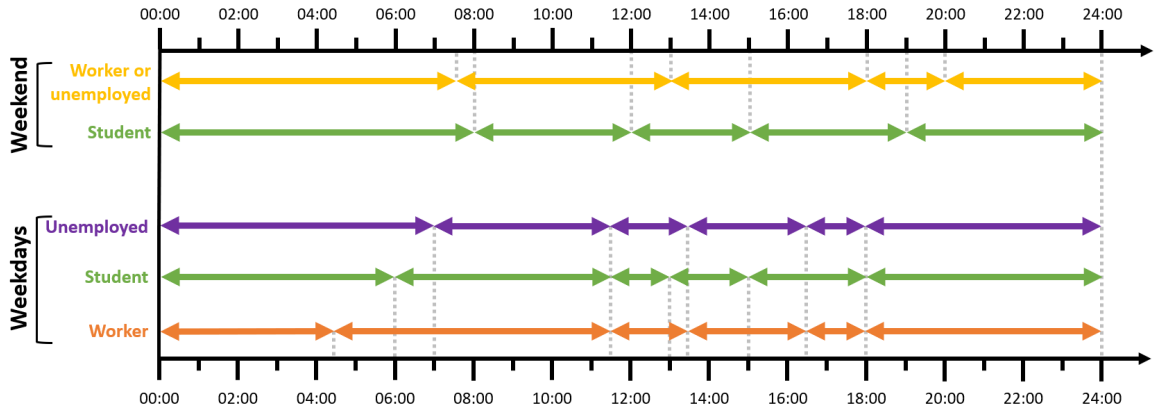


Figure 9: Definition of time slots with method 1 (defined time slots)

For method 2, the way of defining the time slots is in the same idea as for method 1, but in a more precise way. Time slots are defined by finding the relative extrema of the start time distribution for each socioeconomic categories. An example of method 2 can be visualised in Figure 10. The parameters used and the final time slots definition can be seen in Tables 11 and 22 in Appendix.

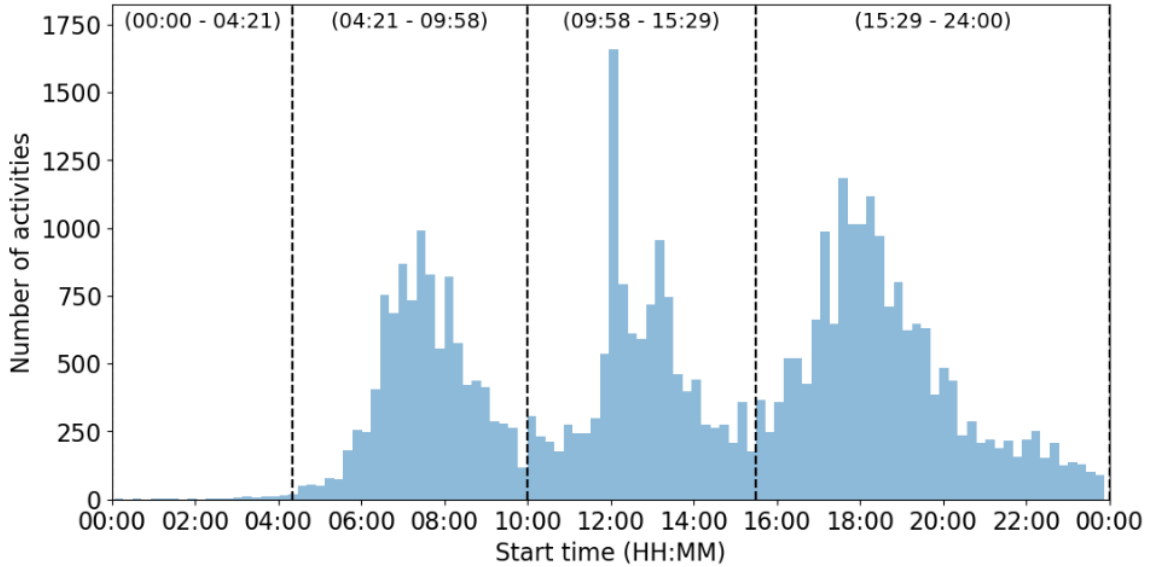


Figure 10: Example of the time slot definition for method 2. Time slot 1 is between 00:00 and 04:21, 2 is between 04:21 and 09:58, 3 is between 09:58 and 15:29, 4 is between 15:29 and 24:00.

Method 3 is to define a number of time slots per 24 hours having the same duration. To decide which number of time slots per day is the most ideal and most likely to capture the preferred activities of individuals, several time slot durations are tested. The variations tested are listed in Table 6. Unlike the first two methods, there is no differentiation here according to employment status or day of the week because we want to see if the duration of the slots can be more adapted to a specific category, and if the quality of the generation depends on the duration of the time slots.

Table 6: Definition of time slots with method 3 (defined number of time slots)

Number of time slots	Duration of each time slot
1440	1 min
288	5 min
144	10 min
96	15 min
48	30 min
24	1 hour
12	2 hours
6	4 hours

Now, we need to test the ability of the three methods described above to capture the preferred activities of the population groups. This is done using the motif comparison method described in Section 4.3. The list of the most popular activities for each socio-demographic categories according to the day of the week are compared with the corresponding motif. Here, all three time slot definition methods are evaluated, using the original data.

The original motifs found with the time slots definition method and the real motif generated with the method described in Section 4.1.1 are presented in Table 7. For method 3, 1-hour time slots are used. The percentages difference between the first and second most common real motifs generated with the methodology of Schultheiss et al. (2021) is shown in brackets.

Table 7: Evaluation of the time slots definition methods

	Method	Weekdays		Week-end	
		Original motif	Real motif	Original motif	Real motif
Full-time worker	1 Defined time slots	H, W, H	H, W, H (15.3%)	H	H, L, H (24.3%)
	2 Automatic time slots	H, W, H	H, W, H (15.3%)	H	H, L, H (24.3%)
	3 Defined number of time slots	H, W, H	H, W, H (15.3%)	H	H, L, H (24.3%)
Part-time worker	1 Defined time slots	H	H, W, H (5.5%)	H	H, L, H (24.1%)
	2 Automatic time slots	H	H, W, H (5.5%)	H	H, L, H (24.1%)
	3 Defined number of time slots	H, W, H	H, W, H (5.5%)	H	H, L, H (24.1%)
Student	1 Defined time slots	H	H, E, H, E, H (2.4%)	H	H, L, H (38.1%)
	2 Automatic time slots	H, E, H	H, E, H, E, H (2.4%)	H	H, L, H (38.1%)
	3 Defined number of time slots	H, E, H, E, H	H, E, H, E, H (2.4%)	H, L, H	H, L, H (38.1%)
Unemployed	1 Defined time slots	H	H, S, H (1.8%)	H	H, L, H (25.4%)
	2 Automatic time slots	H	H, S, H (1.8%)	H	H, L, H (25.4%)
	3 Defined number of time slots	H	H, S, H (1.8%)	H	H, L, H (25.4%)

We start by separating the analysis by day of the week. As far as weekdays are concerned, we can see that all methods manage to capture the most common motif of full-time workers. As we saw earlier in Figure 6, this category of population has clear motifs of activity common to the whole group. Indeed, there is a 15 % difference between the most common motif (H,W,H) and the second most common (H,W,H,W,H). In addition, this motif is very similar to the first one, which also explains the similarity within the three methods.

For part-time workers, only method 3 (defined number of time slots) is able to capture the activity motif. For the other methods, the preferred activity at each time of day is to be only at home. We can see here that the difference between the most common motif (H,W,H) and the second most common (H,L,H) is small (5.5%). This may explain the difficulty of the first two methods in capturing the most common motifs, but the main explanation is that part-time workers do not all follow the same schedule during the day, and have more freedom (e.g., some of them work in the morning, some in the afternoon). Therefore, it is challenging for the method to capture the motifs of this category. For the student category, method 3 is the most successful in capturing activity motifs. Method 2 comes close, but fails to capture the subtlety of going home in between.

Finally, with regard to unemployed people, none of the methods succeeded in capturing the most common motif. This is not surprising, as this is the category with the widest range of activities. Moreover, these activities are performed at random hours. For weekend activities, we can apply the same observations. In fact, the schedules of activities are varied and the rest of the time, the home activity is performed. This is why all three methods have difficulty in capturing activity-based motifs. However, for the student category, method 3 did manage to capture it.

To sum up, method 3 (defined number of time slots) seems to be the best for capturing the preferred activities of a population type at a given time. Since several duration of slots are possible, it is important to identify which one is the most appropriate for generating activities. In the Table 8, the results of the time slots definition of method 3 are presented.

Table 8: Evaluation of the time slots definition for method 3 (defined time slots duration)

	Duration of the time slots	Weekdays		Week-end	
		Original motif	Real motif	Original motif	Real motif
Full-time worker	1/5/10/15/30/60 min	H, W, H	H, W, H (15.3%)	H	H, L, H (24.3%)
	2 hours	H, W, H	H, W, H (15.3%)	H	H, L, H (24.3%)
	4 hours	H, W, H	H, W, H (15.3%)	H	H, L, H (24.3%)
Part-time worker	1/5/10/15/30/60 min	H, W, H	H, W, H (5.5%)	H	H, L, H (24.1%)
	2 hours	H	H, W, H (5.5%)	H	H, L, H (24.1%)
	4 hours	H	H, W, H (5.5%)	H	H, L, H (24.1%)
Student	1/5/10/15/30/60 min	H, E, H, E, H	H, E, H, E, H (2.4%)	H, L, H	H, L, H (38.1%)
	2 hours	H, E, H, E, H	H, E, H, E, H (2.4%)	H, L, H	H, L, H (38.1%)
	4 hours	H	H, E, H, E, H (2.4%)	H	H, L, H (38.1%)
Unemployed	1/5/10/15/30/60 min	H	H, S, H (1.8%)	H	H, L, H (25.4%)
	2 hours	H	H, S, H (1.8%)	H	H, L, H (25.4%)
	4 hours	H	H, S, H (1.8%)	H	H, L, H (25.4%)

When time slots are between 1 and 60 minutes, the results are the same. These ways of defining slots manage to capture weekday activities for full-time and part-time workers as well as students. With 2-hour slots, the main motif of part-time workers is not captured, and with 4-hour slots, only the motif of full-time workers is captured.

As seen above, weekend motifs have difficulty being captured by preferred activity methods. This is due to the diversity of activities, but also to the fact that activity start times are spread out over the day, and have no precise times for the whole population group (see Figure 7). Method 3 and time slots durations from 1 to 120 minutes is therefore only able to capture the most common motif of

students.

We can see that the best way to capture preferred activity is to define slots between 1 and 60 minutes. A higher number of slots implies a larger final dataset and consequently a longer generation time. In addition, for planning studies such as Pougala et al. (2022), a duration of one hour is enough. For these reasons, method 3 with 1-hour time slots is chosen. To have points of comparison, three scenarios are kept for the activities generation and are shown in Table 9.

Table 9: Scenarios for the generation

Scenario	Method
1	Method 3, 10-min time slots
2	Method 3, 1-hour time slots
3	Method 3, 4-hour time slots

5.5 Preferred activity generation

Now that the determining attributes have been identified and the time slots defined, we can generate the list of preferred activities per time slots. In each case, the dataset is separated into weekdays and weekends before the time slots are defined. Then, the conditional tables are formed and generated. Finally, the activities are drawn conditional on the time slot and the employment status.

As a first step, it is worth checking that the generation has succeeded by verifying the attributes used to generate the activities. In Figure 11, the differences between the frequency of the employment statuses generated and the original dataset are shown for the three scenarios.

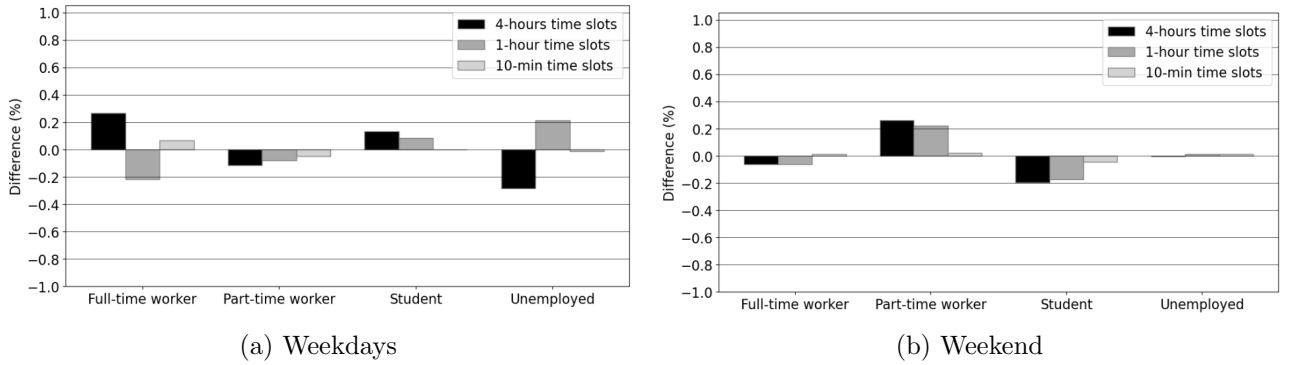


Figure 11: Employment status generated compared to the original dataset, depending on the day of the week

First, we can see that the differences in proportions are all less than 0.3 %, whatever the time slots definition method, day of the week, or employment status. We can see that the difference becomes smaller as the number of slots increases. Note also that there is no noticeable difference in the order of magnitude between the days of the week. In Figure 12, the differences in proportion between the type of activities generated and the original dataset are shown as a function of the time slot definition method.

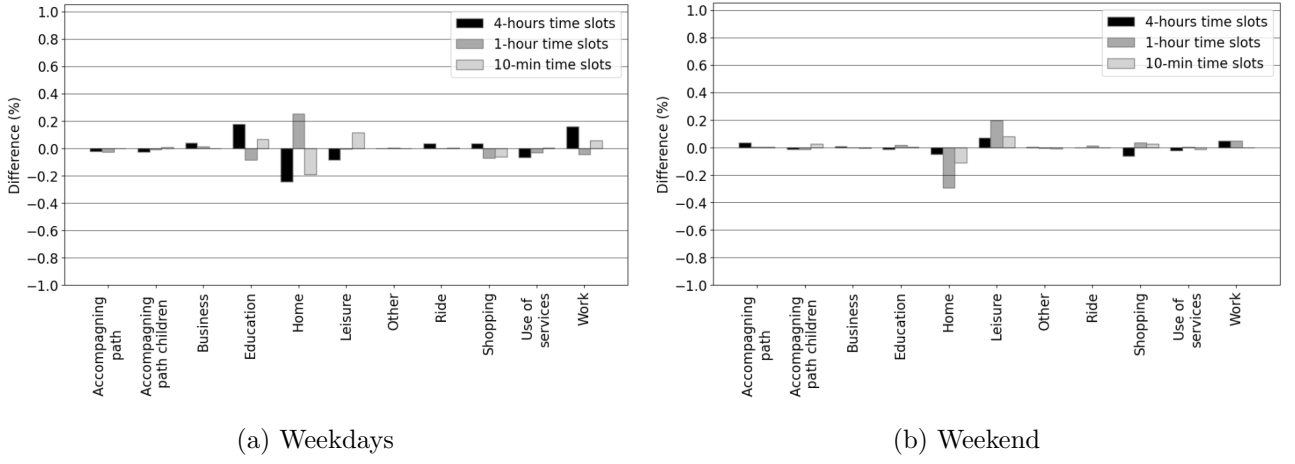


Figure 12: Activity type generated compared to the original dataset depending on the day of the week

The same comments as for employment status generation apply here to the activities generated. Regardless of activity type, definition method, or day of the week, the proportion differences are less than 0.3 %. The previous observation concerning the reduction of differences in the larger number of slots also applies here in the majority of cases. Note also that the activities with the biggest differences are the most practised ones (home, leisure, education, shopping, or work). The differences in proportion of time slots generation and original dataset are shown in Figure 13. The results of the generations with 4-hour and 1-hour time slots are shown.

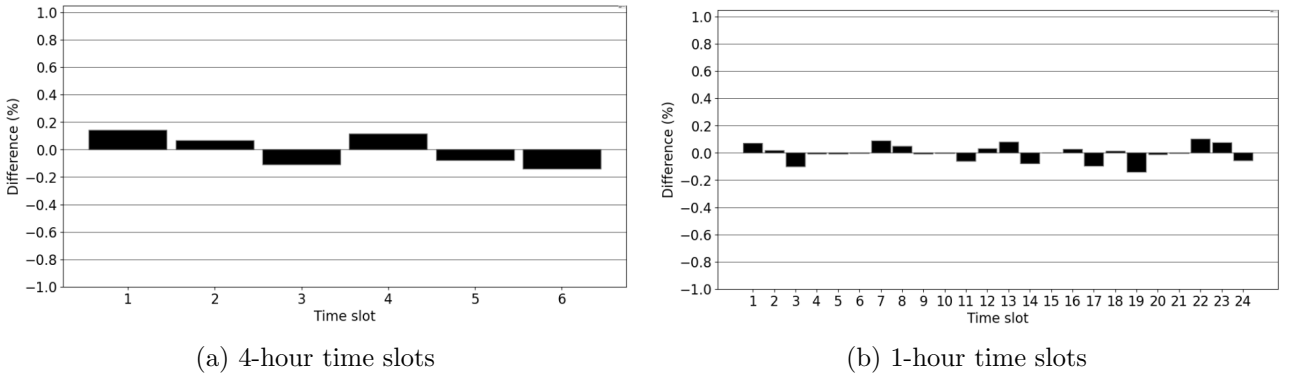


Figure 13: Time slots generated compared to the original dataset depending on the time slot definition

As before, proportion differences never exceed 0.3 %. The differences are slightly greater with the 4-hour time slots than with the 1-hour time slots. We can sum up that the generation correctly replicates the original dataset when we look at employment status, activities, or time slots individually as there are no major discrepancies.

After checking the generation of the three attributes individually, the list of preferred activity per time slots for each population group is selected to be compared and evaluated with the most performed motif (see methodology in Section 4.3). The three scenarios are compared, and the goal is to see if they capture the activities in the same way as with the original data. The original motif, as well as the motif found after the generation are compared with real most performed motif. The results can be seen in Table 10.

Table 10: Evaluation of the time slots definition before and after the generation

	Scenario	Weekdays			Week-end		
		Original motif	Generation motif	Real motif	Original motif	Generation motif	Real motif
Full-time worker	1	H, W, H	H, W, H	H, W, H (15.3%)	H	H	H, L, H (24.3%)
	2	H, W, H	H, W, H	H, W, H (15.3%)	H	H	H, L, H (24.3%)
	3	H, W, H	H, W, H	H, W, H (15.3%)	H	H	H, L, H (24.3%)
Part-time worker	1	H, W, H	H, W, H, W, H, W, H, W, H	H, W, H (5.5%)	H	H	H, L, H (24.1%)
	2	H, W, H	H, W, H	H, W, H (5.5%)	H	H	H, L, H (24.1%)
	3	H	H	H, W, H (5.5%)	H	H	H, L, H (24.1%)
Student	1	H, E, H, E, H	H, E, H, E, H	H, E, H, E, H (2.4%)	H, L, H	H, L, H	H, L, H (38.1%)
	2	H, E, H, E, H	H, E, H, E, H	H, E, H, E, H (2.4%)	H, L, H	H, L, H	H, L, H (38.1%)
	3	H	H, E, H	H, E, H, E, H (2.4%)	H	H	H, L, H (38.1%)
Unemployed	1	H	H	H, S, H (1.8%)	H	H	H, L, H (25.4%)
	2	H	H	H, S, H (1.8%)	H	H	H, L, H (25.4%)
	3	H	H	H, S, H (1.8%)	H	H	H, L, H (25.4%)

We can see that the preferred activities evaluated after the generation (generation motif) are almost the same as those found on original data (original motif). The generation with 10-min time slots struggles to capture the most common motif of part-time workers. However, the generation with 4-hour time slots is more successful in capturing the most common student motif. In each case, we can see that the percentages of the most popular activities are close together, making it difficult to identify the most common motif. On the other hand, it shows that the most performed activities per slot are close and that it only takes a small variation for the method to succeed in capturing the real motif or not. As in the original data study, the best generation method is with 1-hour time slots.

5.5.1 Probabilities of preferred activities

It is now possible to view the generation of the preferred activities by time slot of the day. The data show the probability of each type of activity per time slot. As the generation method with 1-hour time slots proved to be the best, the following results are based on this method. In Figure 14, the probabilities of activities by time slot and employment status for weekdays are presented.

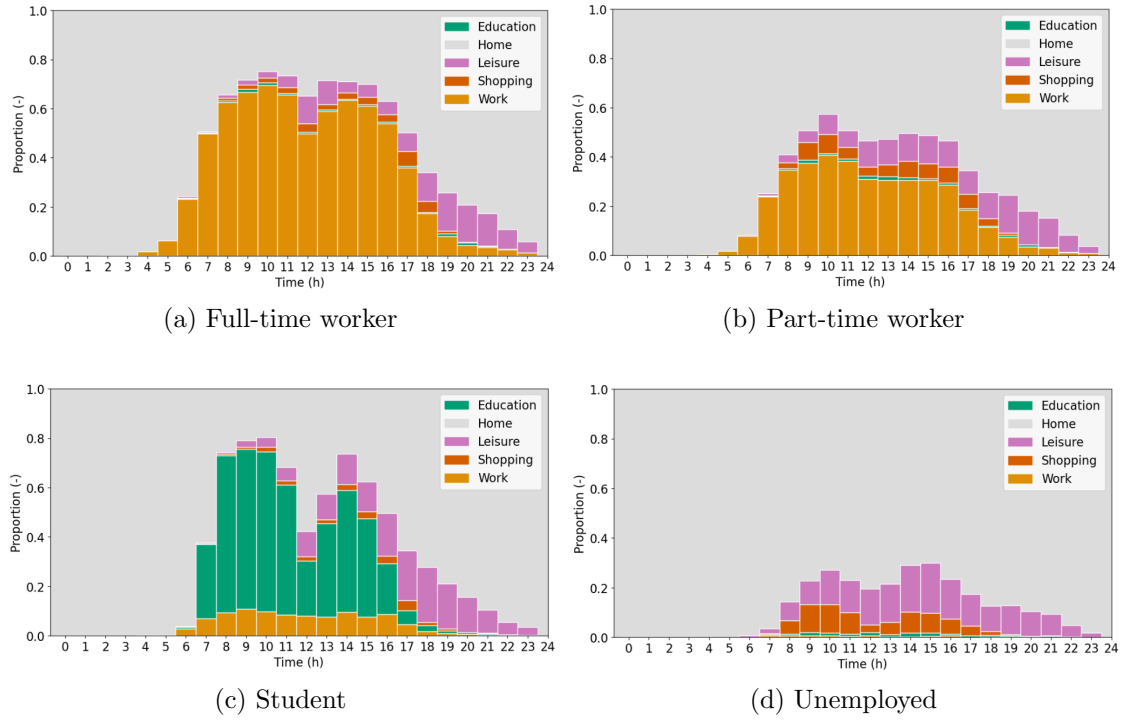


Figure 14: Probabilities of the activities in a specific time slot for the different working statuses during weekdays

We can clearly see that the main activity of full-time workers during the day is as expected work (more than 60 % of the activities during working hours). We can note the drop (10 %) in the probability of this activity around midday (lunch and going home for some people). As the day progresses, the probability of work decreases in favor of leisure. We can see that from 06:00 to 17:00, the majority of people are away from home to perform activities (mainly work or leisure).

In the case of part-time workers, the majority of activities during the work period are related to the work activity. A higher probability of leisure activity is observed from the start of the day and remains stable throughout. In contrast to full-time workers, the probability of people who are at home during working hours remains high. This is not surprising, given the definition of part-time workers.

For people categorized as students, the probability of education activity is high (>70 % during the morning). We can clearly see the class periods and the time around midday when they go home to eat. Leisure activities have a higher probability than for the others activities and extend mainly into the afternoon. There is a base of work activities (<10 %) between 06:00 and 17:00 that probably comes from students working alongside their studies and for whom the day of the micro-census fell on their working day.

Finally, we can see that unemployed people have a high probability of spending their time at home (more than 70 %). In the afternoons, a similar probability as for the students is for leisure activities (one of the reasons is probably to accompany them). Shopping is also an important probability of the unemployed people. We imagine that this category is made up of homemakers looking after the home and family. Even if these people are at home, that does not mean they do not do anything - quite the opposite, in fact. But for the purposes of this study, we are not looking to capture these home-activities. The preferred activity results for the weekend according to the employment status are shown in Figure 15.

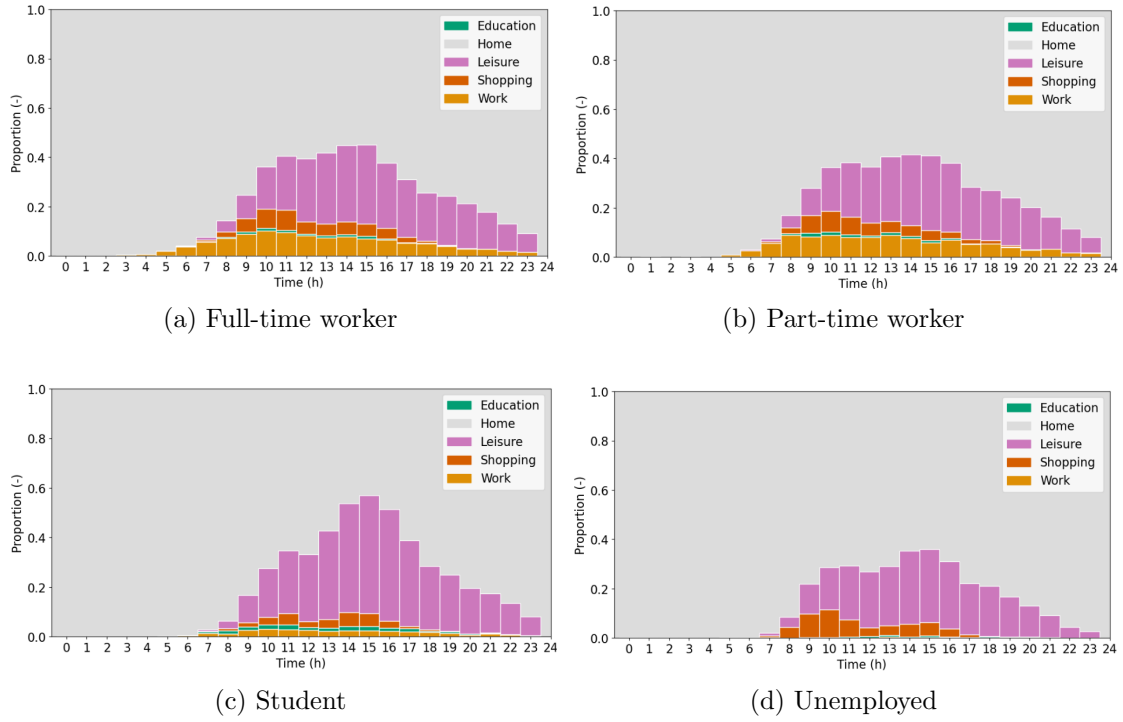


Figure 15: Probabilities of the activities in a specific time slot for the different working statuses during the weekend

As expected, the differences between categories of people are much smaller than on weekdays. We can see that there are nearly no differences between people working full-time or part-time. In both cases, from 09:00, the majority of the time is taken up by leisure activities, if we except home activities. A higher probability than during the weekdays is dedicated to shopping for these two categories. Compared to students and unemployed people, there is a higher probability of work activities throughout the day. These are people who have to work at weekends, in the service sector for example.

Students' activities are distributed differently from the first two categories. In fact, their probability of leisure activities is higher, particularly with a peak around 15:00. Education, shopping, and work are also present, but with smaller probabilities. For unemployed people, the distribution is similar to the employed people, but without the probability of work. Except for home activities, leisure and shopping activities are predominant.

To sum up, we understand the difficulty of the validation method in capturing the most common motifs in each category based on the probabilities of activities most practiced per time slot. Indeed, for the weekend or the unemployed during the weekdays, the majority of time is spent at home. Activities performed outside the home have variable schedules within the population, and it is therefore difficult to see clear trends. Even if the majority of people do not stay at home the whole weekend, not everyone leaves the house at the same time. The only category that has a large enough probability of an activity type to be captured is students, who are more likely to have leisure activities than others.

As with individual attributes, it is important to validate the activity generation according to the time slots by looking at the differences between the original and the generated sample. Figure 16 shows these differences for full-time workers on weekdays. The results for the other people categories are available in Appendix XX.

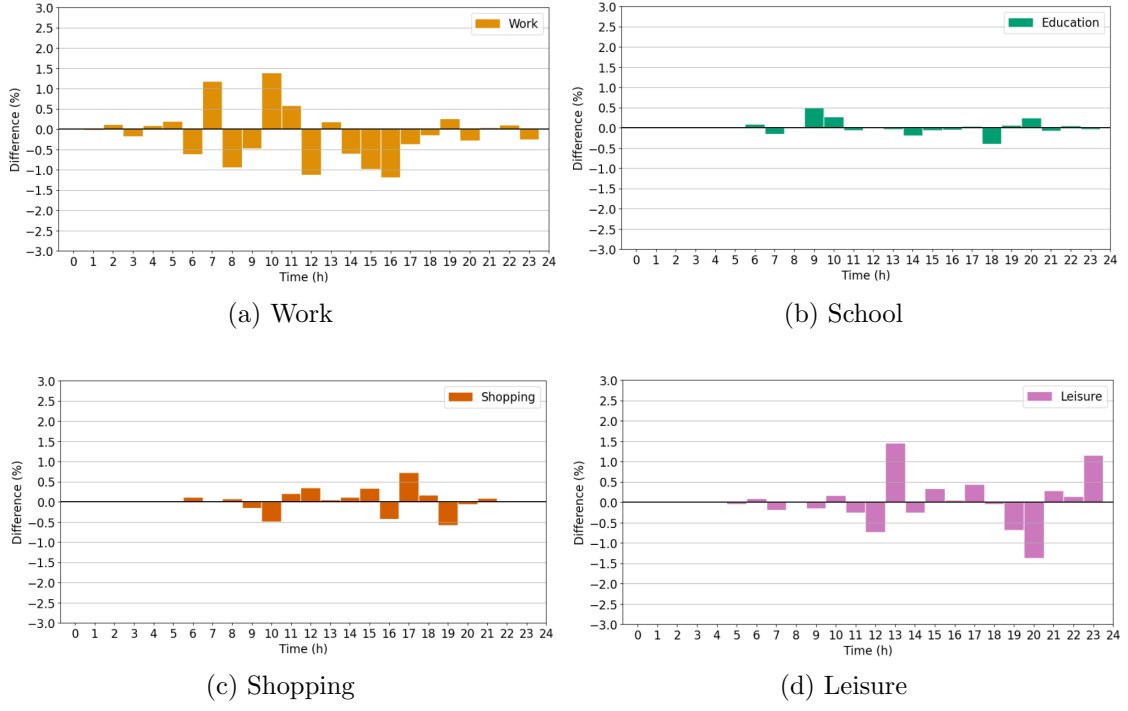


Figure 16: Comparison with real data of the activities generated per time slots on weekdays

We can see that the differences never exceed 1.5 % in absolute terms. It is higher (around a factor of 5) than for activities or slots tested individually, but the result is still satisfying. The probabilities of activities turned out to be close to very close to reality and the results presented here show that the generation is efficient and correctly reproduces the original preferred activities distribution. Work and leisure activities seem to be more affected and to have more differences than school or shopping activities, which are less popular activities for full-time workers. It is interesting to note that during the period when the vast majority of people are at home (early morning), the generation works well, with virtually no differences whatever the activity.

To sum up, we found that the definition using 1-hour time slots provided the best visualization of preferred activities. During generation, this method captured the most common weekday patterns of full-time and part-time workers, as well as students. For weekends, it was more difficult to capture the most common motifs for the reasons mentioned above (disparate schedules, long periods at home). Finally, the probabilities distributions of all preferred activities per time slot were synthetically generated for weekdays as well as weekends, according to people’s employment status. These data enabled us to understand the probabilities of activities over a day for a defined population. These data were synthetically generated with good correspondence to the original data (max 1.5 % deviation).

6 Discussion

Following the results, feature selection (motif analysis, start times and durations) is discussed. Then, the method of how to define time slots and generate preferred activities is analyzed. Finally, the validation method is discussed.

6.1 Feature selection

The goal here is to divide the dataset to identify the important variables. We segment it according to the day of the week and the employment status. This proves to be the right segmentation, providing a good characterization of the population. The main challenge of this work is to capture as much information as possible without losing any. It is always possible to segment the population further, but it is not necessarily a good thing. Generally speaking, we can say that the segmentation and the identification of important variables are essential for the success of future generations.

An important way to identify significant variables is to generate activity-location based motifs. After segmenting the population, we can see clear differences according to employment status or day of the week. The most popular activities and the way in which they are practiced during the day are completely dependent on these variables. This corresponds to what we would expect, but it is interesting to know the proportions of the motifs practiced to generate the activities better. As seen in the results, each category of people has its own characteristics. On weekdays, workers' motifs are significantly linked to work (with a lower proportion for part-timers), while students' motifs are largely and almost exclusively linked to education. Unemployed people have less varied motifs, as they spend more time at home.

The study of these motifs shows that activity preferences of certain categories are more likely to be captured, because they are more pronounced (e.g., full-time workers or students). It should also be noted that, despite the high percentages of the first motifs, and therefore clear preferred activities, the top 10 most common motifs does not necessarily represent a high percentage of the data. Intentionally, the motifs are not grouped together, resulting in a wide variety. In Schultheiss et al. (2021), the 10 most common motifs accounts for 80% of activities. Here, only students and unemployed people on weekends exceed 70%. In contrast, part-time workers on weekdays have only 37% of activities represented in their top 10. Therefore, we can say that it would be necessary to group motifs to have a higher representation of activities in the top 10, with the risk of losing some information.

From the motif analysis, the four most common activities are retained (work, school, shopping, and leisure). In terms of start times, we can see that work and school activities have marked peaks at the times we would expect (beginning or middle of the day). This confirms the idea already mentioned in the motifs analysis that full-time workers and students have clear preferred activities that are easier to capture than other categories. We can also see that weekend activities will be harder to capture, given the higher proportion of leisure and shopping activities that are spread out over the day and have no specific timetable like work or school. In terms of durations, shopping and leisure activities are much shorter than the others, making them more complicated to capture, especially combined with their random timetables.

In addition to the attributes discussed above, other are used to segment the population (general public transport pass holder, hierarchical position at work, self-employed, gender, etc.). These segmentations reveal some differences in the feature distributions, but they are not significant factors in defining time slots or generating preferred activities. Therefore, these classifications are not taken into account. It is important to bear in mind that over-segmentation is not necessarily beneficial, as samples get smaller and generating preferred activity can become more complicated. The ideal is to be able to segment enough to capture the differences, but not too much to avoid losing information.

In the end, the feature selection process enables us to see the important attributes (day of the week and employment status).

6.2 Definition of time slots

Three different methods are used to define time slots. The first two methods based on the specificities of the population categories are not able to capture all the subtleties of preferred activity. Clear motifs (full-time workers and students during the week) are captured, but not the others. In the end, the simplest method of application (method 3, defined time slots duration) best managed to replicate the original distribution without losing any information. Several time slots durations are considered, but 1-hour time slots prove to be the best. They allow conditional tables to be relevant, as there is enough data per time slot, while being numerous enough to capture the information. Scheduling studies such as Pougala et al. (2022) are satisfied with preferred activity probabilities on 1-hour time slots. According to the most common motif validation method, the 1-hour time slots method captures the maximum preferred activities, while having the highest time slots duration.

6.3 Preferred activity generation

Following the definition of time slots, three scenarios from method 3 are selected for the generation. The validation method of the most common motif confirms that 1-hour time slots are the most suitable for generating preferred activities. The final result shows the probabilities of activities performed in 1-hour time slots for each population group, according to employment status (full-time, part-time, student, or unemployed) and day of the week (weekdays or weekends). We can clearly see the differences between the groups, and it is possible to use this data for other studies, such as scheduling for example. This data is synthetic and is generated conditional to employment status and time slots in the day. The fact that the data is synthetic means that privacy issues can be avoided, and the size of the sample generated can be chosen. Various comparisons with the original distribution shows that the dataset is correctly generated and replicated. This confirms that the attributes chosen during the feature selection phase are relevant and that this work succeeds to provide a list of activities probabilities that can be use by other studies.

One of the challenges of this project is to successfully validate the generated data. The method used has satisfying results in the comparison between the distributions of the original and generated datasets. To check whether the method applied successfully captures the preferred activities of population groups, for each individual we choose the activity with the highest probability per time slot, and create a motif out of these activities. Then, we compare frequencies of synthetic motifs with real ones. The results show that this method works well when the activities practiced represent a majority of activities compared to the Home activity. When this activity is too preponderant, or when the other activities have similar distributions, it is more complicated to confirm the generation by the motif comparison method. The limitation of this project is therefore the struggle to have a clear method to confirm which motifs are most practiced on a given day by a category of individuals. However, the data generated correctly replicates the original distribution and provides the probabilities of each activity practiced per time slot.

7 Summary and conclusion

The aim of this research was to generate activity-related features, avoiding the constraints associated with costly and time-consuming surveys. The synthetic data were to be usable for future studies, particularly in the field of schedule generation. By generating synthetic data, researchers can create customized datasets tailored to their specific needs, giving them total control over variables and activity attributes. By providing realistic synthetic data, this research contributes to improving the accuracy and relevance of scheduling models.

To achieve this objective, the Gibbs sampler algorithm was considered as a method for generating such synthetic data. It is well suited to dealing with discrete and sequential variables, making it an appropriate choice for generating activities with attributes such as start times, durations and activity types. Using the Gibbs sampler algorithm, this research aimed to generate a list of preferred activity by replicating the distribution and characteristics of the real sample. In this work, we addressed two main research questions. First, how to represent and capture a population's preferred activities, including start times and duration? And second, how to synthetically generate these preferred activities using the Gibbs sampler algorithm?

To answer the first question, we first carried out a preliminary study on real data to see which attributes have an impact on activity-location based motifs, activity type, start times and duration. Our work showed that attributes such as the day on which the activity is performed or the individual's employment status have a significant impact on the distribution of preferred activities. The characteristics of each population group were highlighted and used to define how to generate synthetic preferred activities. The preliminary study performed in this project has shown how useful and interesting it is to segment the population. It is important for the generation of synthetic data that the groups of people represent the population in a characteristic way to be relevant.

With regard to the second research question, we used the Gibbs sampler algorithm to generate preferred activities synthetically. The preliminary study helped us to define how to generate the data and capture timing attributes related to the activities. This was done by segmenting the day into time slots. Several methods were tested, but the study showed that defining slots of equal duration of one hour was the best for capturing preferred activities. The results obtained with the generation were satisfying and the synthetic values reproduced the original values almost perfectly, indicating that the Gibbs sampler algorithm was capable of capturing activity variations and preferences.

It should also be stressed that validation of the generated data is a crucial step in guaranteeing its reliability and relevance. In this work, we performed validation by comparing the generated activity distributions with the original data and the results were very encouraging. Moreover, the most common motifs in each generated population group were able to be compared, and confirmed the generation with the real data in the majority of cases. Where this was not possible, the study was able to show the reasons for this (too much diversity of activity, too much time at home, undefined schedules, etc.). However, a more rigorous approach using appropriate statistical validation methods would be required to boost confidence in the data generated. One way of addressing this issue is to be able to test synthetic data directly in a scheduling model. This is precisely why the data is generated and if the scheduling model manages to generate results of the same quality as with real data, this validates the generation.

The results of this research have important implications for the field of synthetic activity-related features generation. One of the main contributions of this research lies in the use of the Gibbs sampler algorithm to generate synthetic activity preferences. Using this algorithm, we are able to take into account the attributes of activities, which is crucial for correctly modeling activity behaviors. By proposing a method for generating synthetic data, this study offers an economical and practical alter-

native for obtaining relevant data without the need for costly and time-consuming surveys. Through the generation of synthetic data, larger datasets can be generated, enabling a better understanding of activity behavior. The generation of list of preferred activities is a data that is not present in the original dataset and that is useful for other study. By almost perfectly reproducing the original value distributions, this generation method offers high accuracy in the representation of preferences and activity characteristics. The synthetic data generated can be used in a variety of practical applications, such as simulating scheduling scenarios, optimizing timetables, forecasting travel demand and evaluating mobility policies. These applications can help decision-makers better understand user needs, assess the impact of policy changes and make informed decisions to improve transportation services.

Finally, this work enables us to generate synthetic preferred activities according to time of day and population group. This would allow them to dispense with real data, and use synthetic data instead. Although this project shows that data generation reproduced the original data, more rigorous validation methods are recommended to assess the quality of the synthetic data generated. This may involve the use of appropriate statistical metrics, to go further than the distribution or comparison analyses with the most common motifs, done in this work to compare the generated data with real data. Further validation of the generated data will ensure their relevance and usefulness for practical applications. It is important to stress that the generation of synthetic data does not completely replace the use of real data. Real data remains essential for understanding the specific contexts, individual behaviors and social factors that can have an impact on scheduling. Synthetic data should be used as a complement to real data, providing additional information and enabling scenarios to be explored.

References

- Andrade, C. (2020). The Limitations of Online Surveys. *Indian Journal of Psychological Medicine*, 42(6), 575–576. <https://doi.org/10.1177/0253717620957496>
- Arentze, T., Timmermans, H., & Hofman, F. (2007). Creating synthetic household populations: Problems and approach. *Transportation Research Record*, (6), 85–91. <https://doi.org/10.3141/2014-11>
- Axhausen, K. (2000). Activity-based modelling research directions and possibilities. *Arbeitsberichte Verkehrs- und Raumplanung*, 48. <https://doi.org/10.3929/ethz-a-004241843>
- Bae, K.-H. G., Feng, B., Kim, S., Lazarova-Molnar, S., Zheng, Z., Roeder, T. M. K., & Thiesing, R. M. (2020). Synthetic trip list generation for large simulations. <https://informs-sim.org/wsc20papers/061.pdf>
- Bayes, A. (2012). The Traditional Four Steps Transportation Modeling Using Simplified Transport Network: A Case Study of Dhaka City, Bangladesh. *International Journal of Advanced Scientific Engineering and Technological Research*, 1, 19–40.
- Beckman, R., Baggerly, K., & McKay, M. (1996). Creating synthetic baseline populations. *Transportation Research Part A: Policy and Practice*, 30(6), 415–429. [https://doi.org/10.1016/0965-8564\(96\)00004-3](https://doi.org/10.1016/0965-8564(96)00004-3)
- Berke, A., Doorley, R., Larson, K., & Moro, E. (2022). Generating Synthetic Mobility Data for a Realistic Population with RNNs to Improve Utility and Privacy. *Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing*, 964–967. <https://doi.org/10.1145/3477314.3507230>
- Bierlaire, M., Ben-Akiva, M., McFadden, D., & Walker, J. (2021). Discrete choice analysis, 430–434.
- Borysov, S., Rich, J., & Pereira, F. (2019). How to generate micro-agents? A deep generative modeling approach to population synthesis. *Transportation Research Part C: Emerging Technologies*, 106, 73–97. <https://doi.org/10.1016/j.trc.2019.07.006>
- Cao, J., Li, Q., Tu, W., & Wang, F. (2019). Characterizing preferred motif choices and distance impacts. *Jinjun Tang, Central South University, CHINA*. <https://doi.org/10.1371/journal.pone.0215242>
- Chu, Z., Cheng, L., & Chen, H. (2012). A Review of Activity-Based Travel Demand Modeling. *Conference: The Twelfth COTA International Conference of Transportation Professionals*. <http://dx.doi.org/10.1061/9780784412442.006>
- Damm, D., & Lerman, S. R. (1981). A theory of activity scheduling behavior. *Environment and Planning A*, 13, 703–718. <https://journals.sagepub.com/doi/pdf/10.1068/a130703>
- Ectors, W., Kochan, B., Janssens, D., Bellemans, T., & Wets, G. (2019). Exploratory analysis of Zipf’s universal power law in activity schedules. *Transportation*, 46, 1689–1712. <https://doi.org/10.1007/s11116-018-9864-9>
- Ectors, W., Kochan, B., Janssens, D., Bellemans, T., & Wets, G. (2022). Activity Sequence Generation Using Universal Mobility Patterns. *Transportation Research Record*, 2676(4), 538–553. <https://journals.sagepub.com/doi/10.1177/03611981211062483>
- Farooq, B., Bierlaire, M., Hurtubia, R., & Flötteröd, G. (2013). Simulation based population synthesis. *Transportation Research Part B: Methodological*, 58, 243–263. <https://doi.org/10.1016/j.trb.2013.09.012>

- Federal Office for Spatial Development ARE. (2015). *Mobility and transport microcensus*. Retrieved June 12, 2023, from <https://www.are.admin.ch/are/en/home/mobility/data/mtmc.html>
- Felbermair, S., Lammer, F., Trausinger-Binder, E., & Hebenstreit, C. (2020). Generating synthetic population with activity chains as agent-based model input using statistical raster census data. *Procedia Computer Science*, 170, 273–280. <https://doi.org/10.1016/j.procs.2020.03.040>
- G.O. Roberts, A. S. (1994). Simple conditions for the convergence of the Gibbs sampler and Metropolis-Hastings algorithms. *Stochastic Processes and their Applications*, 49(2), 207–216. <https://www.sciencedirect.com/science/article/pii/0304414994901341>
- Hörl, S., & Balac, M. (2021). Synthetic population and travel demand for Paris and Île-de-France based on open and publicly available data. *Transportation Research Part C*, 130. <https://doi.org/10.1016/j.trc.2021.103291>
- Kitamura, R., Chen, C., & Pendyala, R. M. (1997). Generation of Synthetic Daily Activity-Travel Patterns. *Transportation Research Record*, 1607(1), 154–162. <https://doi.org/10.3141/1607-21>
- Kukic, M., & Bierlaire, M. (2021). Population synthesis at the level of households. *Proceedings of the 21st Swiss Transport Research Conference (STRC), 12-14 September*. Ascona, Switzerland.
- Kukic, M., & Bierlaire, M. (2022). One-step simulator for synthetic household generation. *Proceedings of the 22nd Swiss Transport Research Conference (STRC), 18-20 May*. Ascona, Switzerland.
- Kukic, M., & Bierlaire, M. (2023). Hybrid simulator for capturing dynamics of synthetic population. *Proceedings of the 23rd Swiss Transport Research Conference (STRC), 10-12 May*. Ascona, Switzerland.
- Ma, L., & Srinivasan, S. (2015). Synthetic Population Generation with Multilevel Controls: A Fitness-Based Synthesis Approach and Validations. *Computer-Aided Civil and Infrastructure Engineering*, 30, 152–160. <https://doi.org/10.1111/mice.12085>
- Müller, K., & Axhausen, K. (2010). Population synthesis for microsimulation: State of the art. *STRC 2010*. <https://www.strc.ch/2010/Mueller.pdf>
- Pendyala, R., Bhat, C. R., Goulias, K. G., Paleti, R., Konduri, K., Sidharthan, R., & Christian, K. P. (2012). Simagent population synthesis. *Southern California Association of Governments*. https://www.caee.utexas.edu/prof/Bhat/REPORTS/Simagent_Final_report_3-Popgen_CEMSELTS.pdf
- Pougala, J., Hillel, T., & Bierlaire, M. (2022). Capturing trade-offs between daily scheduling choices. *Journal of Choice Modelling*, 43. <https://doi.org/10.1016/j.jocm.2022.100354>
- Rich, J., Flötteröd, G., Garrido, S., & Pereira, F. (2021). Review of population synthesis methodologies. *DTU Transport, Technical University of Denmark*.
- Schneider, C., Belik, V., Couronné, T., Smoreda, Z., & González, M. C. (2013). Unravelling daily human mobility motifs. *Journal of the Royal Society Interface*. <http://dx.doi.org/10.1098/rsif.2013.0246>
- Schultheiss, M.-E., Pougala, J., & Kukic, M. (2021). Multi-day mobility motifs: Method and applications. <http://infoscience.epfl.ch/record/297185>
- Stopczynski, A., Pietri, R., Pentland, A. '., Lazer, D., & Lehmann, S. (2014). Privacy in Sensor-Driven Human Data Collection: A Guide for Practitioners. *ArXiv, abs/1403.5299*.

- Su, R., McBride, E., & Goulias, K. (2020). Pattern recognition of daily activity patterns using human mobility motifs and sequence analysis. *Transportation Research Part C: Emerging Technologies*, 120. <https://doi.org/10.1016/j.trc.2020.102796>
- Sun, L., Erath, A., & Cai, M. (2018). A hierarchical mixture modeling framework for population synthesis. *Transportation Research Part B: Methodological*, 114, 199–212. <https://doi.org/10.1016/j.trb.2018.06.002>
- Vieira, A. (2020). *Generating synthetic sequential data using gans*. Retrieved January 9, 2023, from <https://pub.towardsai.net/generating-synthetic-sequential-data-using-gans-a1d67a7752ac>
- Yaméogo, B., Gastineau, P., Hankach, P., & Vandanjon, P.-O. (2021). Comparing Methods for Generating a Two-Layered Synthetic Population. *Transportation Research Record*, 2675(1), 136–147. <https://doi.org/10.1177/0361198120964734>
- Ye, X., Konduri, K., Pendyala, R. M., Sana, B., & Waddell, P. (2009). A methodology to match distributions of both household and person attributes in the generation of synthetic populations. *88th Annual Meeting of the Transportation Research Board, Washington, D.C.*

Appendices

A Results

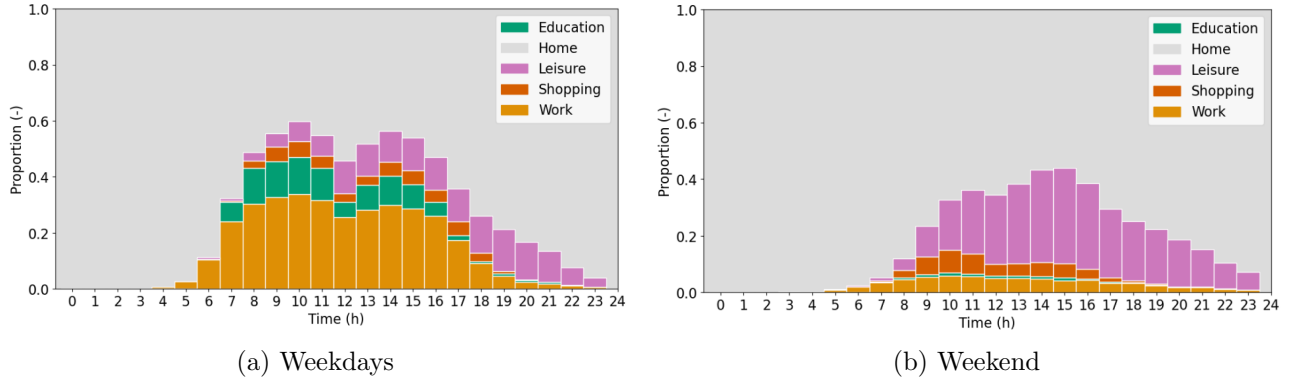


Figure 17: Probabilities of the activities in a specific time slot for the whole population for 1-hour time slots

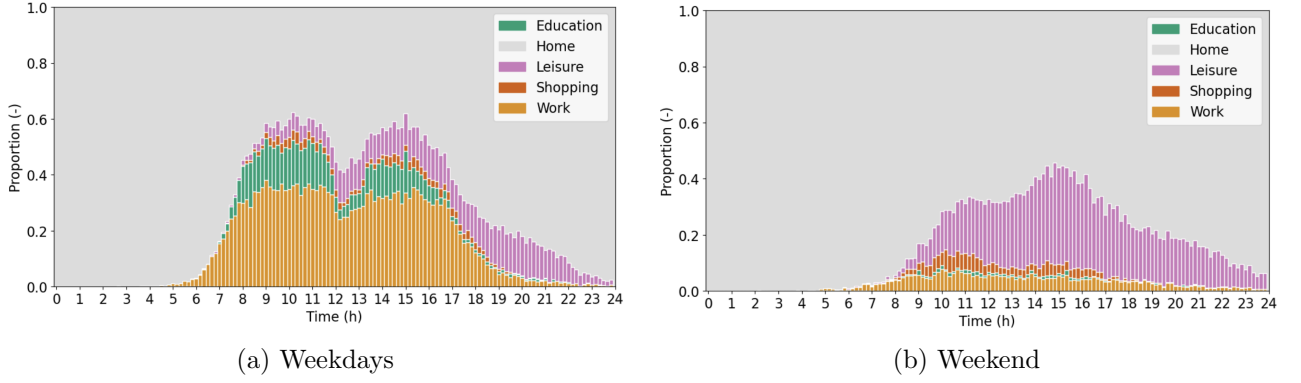


Figure 18: Probabilities of the activities in a specific time slot for the whole population for 10-min time slots



Figure 19: Probabilities of the activities in a specific time slot for the different working statuses during the weekdays for 10-min time slots

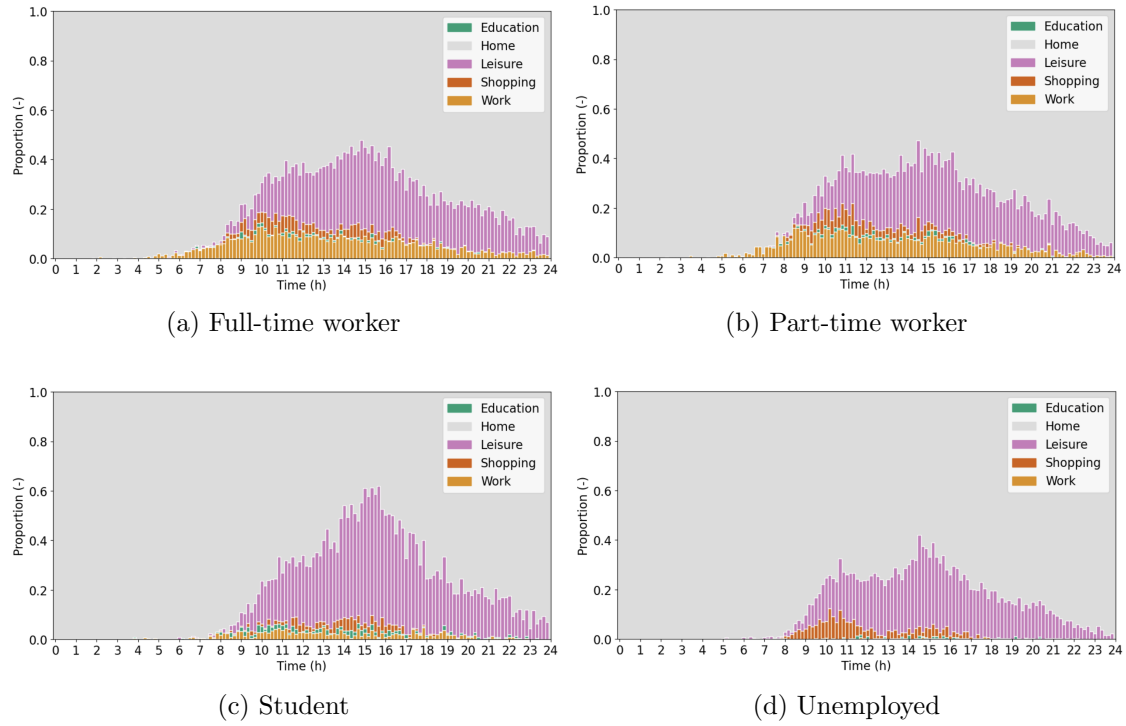


Figure 20: Probabilities of the activities in a specific time slot for the different working statuses during the weekend for 10-min time slots

B Time slots definition and parameters

Full-time worker			Part-time worker		
Slot	Start	End	Slot	Start	End
1	00:00	05:00	1	00:00	05:00
2	05:00	11:30	2	05:00	11:30
3	11:30	13:30	3	11:30	13:30
4	13:30	16:30	4	13:30	16:30
5	16:30	18:00	5	16:30	18:00
6	18:00	00:00	6	18:00	00:00

Student			Unemployed		
Slot	Start	End	Slot	Start	End
1	00:00	06:00	1	00:00	07:00
2	06:00	11:30	2	07:00	11:30
3	11:30	13:00	3	11:30	13:30
4	13:00	15:00	4	13:30	16:30
5	15:00	18:00	5	16:30	18:00
6	18:00	00:00	6	18:00	00:00

(a) Weekdays

Full-time worker			Part-time worker		
Slot	Start	End	Slot	Start	End
1	00:00	07:30	1	00:00	07:30
2	07:30	13:00	2	07:30	13:00
3	13:00	18:00	3	13:00	18:00
4	18:00	20:00	4	18:00	20:00
5	20:00	00:00	5	20:00	00:00

Student			Unemployed		
Slot	Start	End	Slot	Start	End
1	00:00	08:00	1	00:00	07:30
2	08:00	12:00	2	07:30	13:00
3	12:00	15:00	3	13:00	18:00
4	15:00	19:00	4	18:00	20:00
5	19:00	00:00	5	20:00	00:00

(b) Weekend

Figure 21: Definition of time slots with method 1 (defined time slots)

Full-time worker			Part-time worker		
Slot	Start	End	Slot	Start	End
1	00:00	04:21	1	00:00	05:36
2	04:21	09:58	2	05:36	10:28
3	09:58	15:29	3	10:28	16:28
4	15:29	00:00	4	16:28	00:00

Student			Unemployed		
Slot	Start	End	Slot	Start	End
1	00:00	06:18	1	00:00	07:05
2	06:18	10:31	2	07:05	12:45
3	10:31	14:28	3	12:45	20:59
4	14:28	23:29	4	20:59	23:29
5	23:29	00:00	5	23:29	00:00

(a) Weekdays

Full-time worker			Part-time worker		
Slot	Start	End	Slot	Start	End
1	00:00	07:52	1	00:00	07:55
2	07:52	12:59	2	07:55	12:44
3	12:59	00:00	3	12:44	00:00

Student			Unemployed		
Slot	Start	End	Slot	Start	End
1	00:00	09:51	1	00:00	07:54
2	09:51	14:57	2	07:54	13:58
3	14:57	22:56	3	13:58	23:28
4	22:56	00:00	4	23:28	00:00

(b) Weekend

Figure 22: Definition of time slots with method 2 (automatic time slots definition)

The specific parameters for the sensitivity of the two methods for method 2 are described in Table 11 below. The parameters are chosen by testing them with the original data dataset and observing the quality of the time slot definition. Bins are the bars in the histogram. The greater their number, the shorter the interval covered by a bin. Conversely, the fewer the bins, the wider the interval. It is important to find the right balance to be able to see details, but not too much to capture general trends. The order corresponds to a specific parameter of the `argrelextrema` method. The higher the order, the more the method will find the dips. In fact, the method works by looking at the difference in values between the bin neighbors of the analyzed bar and determines whether it is a trough or not according to the sensitivity given by the order. Finally, the threshold provides the sensitivity for calculating the start of the slope, which gives the first interval.

Table 11: Parameters of method 2

Parameter	Value
Bins	108
Order	10
Threshold	1