Open data travel demand synthesis for agent-based transport simulation

A case study of Paris and Île-de-France

Working Paper

Author(s):
Hörl, Sebastian; Balać, Miloš

Publication date:
2020-11-27

Permanent link:
https://doi.org/10.3929/ethz-b-000412979

Originally published in:
Arbeitsberichte Verkehrs- und Raumplanung 1499

Open data travel demand synthesis for agent-based transport simulation: A case study of Paris and Île-de-France

Sebastian Hörl[a,b,*], Milos Balac[a]

[a] *Institute for Transport Planning and Systems, ETH Zurich*
[b] *Institut de Recherche Technologique SystemX, Palaiseau 91120, France*

Abstract

Synthetic populations of travellers and their detailed mobility behaviour are an important basis for agent-based transport simulations, which are increasingly used in transport planning and research today. While previous applications of such simulations put focus on analyzing policies and novel modes of transport, less importance is put on the aspects of data collection, processing and validation to create the disaggregate travel demand. This paper proposes an open-source and extendable pipeline for travel demand synthesis for Île-de-France, which establishes a clear path from raw data to the final demand consisting of households, persons, and trips with their respective attributes. The travel demand synthesis is based on open and publicly accessible data and can be easily transferable to other regions in France. By basing the process on open software and open data, the paper establishes a baseline not only for generating a fully replicable travel demand, but also enables researchers to perform replicable transport simulations. While the proposed pipeline, which is runnable by any reader, is based on rather straightforward and data-driven algorithms, pathways are provided for extending the process. Additionally, methods for quality assessment of synthetic populations and travel demand are proposed and discussed.

Keywords: open, agent-based, transport, simulation, synthetic, population, Paris, Île-de-France

1. Introduction

Agent-based models have become popular in recent decades in many fields of science, mostly because they can model complex systems and interactions within. Importantly, they also allow the modeling of emergent behavior (Bonabeau, 2002). In transportation, traditionally, aggregated four-step models are used to evaluate new policies or infrastructure investments. However, these models fail to capture the interactions between individuals and model transportation on an aggregated scale. They overlook the importance of individuals, their interaction, decisions, and behavior.

Furthermore, they are not readily adaptable to deal with new mobility solutions like car-sharing, bike-sharing, micro-mobility, inter-modality, or future mobility solutions and their operational challenges. It is then no wonder that agent-based models have also found an application in transportation science. Unfortunately, unlike four-step models, agent-based models are challenging to build and maintain, and are usually very data-hungry.

The foundation of every agent-based model in transportation is a synthetic travel demand including a synthetic population of households and persons with sociodemographic attributes and their daily activity patterns in time and space. While there are efforts to document the process of obtaining the synthetic travel demand, those are rarely reproducible, easily verifiable, or open-source. Therefore, they lack one or more of the following: the possibility to be validated by others; the possibility to be extended by others; ease of adding new data or features; and access to the methods used and their documentation.

This paper builds upon the thinking that scientific work should be easily accessible, open-source, and reproducible. Therefore, we propose a fully automated and customizable open-source pipeline for travel demand synthesis. The pipeline takes the raw data and, through various steps, produces a synthetic population. In this way, it allows us to reliably reproduce the required input data for many agent-based transport simulation frameworks. It reduces the effort of building and maintaining the synthetic travel demand.

*Corresponding author
Email addresses: sebastian.horl@irt-systemx.fr (Sebastian Hörl), milos.balac@ivt.baug.ethz.ch (Milos Balac)

Furthermore, for the first time, with this approach, it is possible to perform sensitivity analysis, not only on a static input population, but on the whole process of setting up a transport model from data processing to final simulation. The pipeline itself establishes a travel demand synthesis process based on straightforward data-driven algorithms. This way, it is a comprehensive benchmark for more advanced and novel algorithms, be it in travel demand synthesis (Bayesian networks, Markov models, and others), mode choice models, location assignment, or other components.

2. Background

A long-established approach to forecast transportation demand is to use four-step models. These aggregated models are a traditional way of evaluating policies for large infrastructure investments, and focus on large car and transit flows. Activity-based models emerged to overcome the aggregation drawback of these models. See Recker (1995); Axhausen and Gärling (1992); Kitamura (1988) for early reviews of activity-based models and Rasouli and Timmermans (2014) for a more recent overview. They use various methods to schedule activities for individuals and make mode or destination choices, within a single framework. These models were the answer to the aggregation drawback of four-step models and the inability to model tour constraints.

Moreover, activity-based models were able to provide policy implications on many more dimensions than four-step models. However, these models usually involve a range of econometric sub-models that need to be estimated, and later calibrated to fit the data. Unfortunately, many activity-based models only focus on a small number of regions, are not easily extendable, not open-source, or lack documentation to achieve reproducibility of studies. Examples of activity-based models are CEMDAP (Bhat et al., 2008), which is based on the Dallas-Fort Worth region in the USA, or the rule-based model TASHA (Hao, 2009), which is specifically designed for the Greater Toronto area in Canada. Another notable example of activity-based models is ActivitySim (ActivitySim, 2020). It is being developed as an open-source platform for activity-based travel modeling by multiple transportation agencies in the USA. Another framework that was developed through the same collaboration is PopulationSim. It creates a synthetic population that contains only the socio-demogrpahic attributes and household structures, but no travel demand, based on the marginal data available from the USA census, which creates the basis for ActivitySim.

Agent-based models appeared as the need to model interactions between individuals became important. Today, this need becomes evident as many different transportation services co-exist, and they are used both in a competing and in a complementary fashion. Often, these forms of transport are highly dynamic as vehicles are managed on a minute-by-minute or second-by-second basis and therefore require modeling on a shorter time-scale than activity-based models usually provide. Some examples of agent-based transport models are POLARIS (Auld et al., 2016), SimMobility (Adnan et al., 2016), SUMO (Lopez et al., 2018), or MATSim (Horni et al., 2016).

Some attempts to pair activity-based models with agent-based models exist. The advantage of activity-based models in this configuration is that the former are well-established and based on sophisticated econometric models, often fed with years of available data sets. The latter can provide fine-grained spatial dynamics and emergent congestion patterns, originating from the detailed simulation of agent interactions. Combinations of activity- and agent-based models have, for instance, been performed for MATSim: Ziemke et al. (2015) apply CEMDAP to generate daily activity patterns for a model of the Berlin area, and Hao (2009) and Diogu (2019) pair MATSim with TASHA's Toronto model. POLARIS makes use of the activity-based model ADAPTS (Auld and Mohammadian, 2009).

Agent-based models are dependent on a synthetic travel demand as input data. The aim is then to simulate how agents behave on the transportation network and how they compete for the infrastructure. Viegas and Martínez (2010), for instance, create a synthetic travel demand using a mobility survey. They use statistical approaches and create an agent population for the region of Lisbon, Portugal, which they later use in different agent-based studies (Martinez et al., 2015; Martinez and Viegas, 2017). In the ecosystem of MATSim, various synthetic populations exist, such as the models for Singapore (Erath et al., 2012) or Switzerland (Bösch et al., 2016; Hörl et al., 2019). While the approaches outlined below in this paper draw from the latter reference, here we propose a reproducible, open-source, and open data approach, contrary to the models of Singapore and Switzerland, which have restricted shareability due to the proprietary nature of the underlying data sets. However, open data models exist for MATSim, such as the Open Berlin model (Ziemke et al., 2019b), Ruhr region in Germany (Ziemke et al., 2019a), or the older model of Santiago de Chile (Kickhofer et al., 2016), which can be named as the oldest open

data model in the MATSim ecosystem[1]. Based on publicly available data, Kamel et al. (2018) propose a first MATSim model including an open-source synthesis tool of the Île-de-France region in France, which subsequently was transferred to case studies on the city of Rouen (Vosooghi et al., 2019a,b). Thus, open-data-based models exist, yet they are only documented as part of a more applied, larger-scope case study, whereas the details of the synthesis process are only described briefly in most cases. This arguably limits the reproducibility of not only those models, but also of the studies conducted with them.

Reproducibility has recently become an important topic in research in many fields. Articles emphasize the lack of information or data for other researchers to be able to replicate studies published in scientific journals (Goodman et al., 2016; Chen et al., 2019; Stark, 2018; Boulton, 2016; Baker, 2016). While the necessary steps to ensure replicability of scientific findings can vary across the scientific disciplines, the researchers generally agree that in order to ensure reproducibility of research outcomes, the authors of a scientific study need to:

- provide access to the raw data used in the study or to provide enough details about their data collection methods
- provide access to the code/tools with which the data was processed before it was used in the study
- provide access to the software used in the study
- ensure that the set-up of the study is either explained in enough detail or it is provided as open-source (in case the study requires implementation of a computer program), in order to ensure accessibility

National science foundations are also emphasizing the need for reproducibility in research. The U.S. National Science Foundation states (Cacioppo et al., 2015): "reproducibility refers to the ability of a researcher to duplicate the results of a prior study using the same materials as were used by the original investigator. That is, a second researcher might use the same raw data to build the same analysis files and implement the same statistical analysis in an attempt to yield the same result [...] Reproducibility is a minimum necessary condition for a finding to be believable and informative". The Swiss National Science Foundation (SNSF) requires, to ensure transparency and reproducibility of research findings, that all data used in the publications funded by SNSF is made publicly available, as long as it meets ethical standards.

Based on this, we want to ensure reproducibility of the travel demand generated in this study by (1) publishing the software developed as open-source, (2) using only publicly available data, (3) documenting the processing methods in detail, and (4) ensuring that the complete process is accessible. While first three points are easily achievable, accessibility is not. In order to ensure accessibility for the wide range of users, to regenerate or replicate the demand presented in this paper, we provide easy to follow instructions through the published documentation of the approach. To replicate the study presented here requires no previous programming skills or access to any commercial software. Even though the pipeline does not work as a GUI executable the effort needed to prepare it to run is minimal. Therefore, almost any researcher with access to internet can perform the presented study and obtain the same results. While accessibility has been defined in a wider sense (e.g. Lovelace, 2020) including all potential users from the public, we believe that our framework is a valuable development toward that goal.

Furthermore, the paper at hand adds to the discussion in a number of ways. We provide:

- an integrated, open-source software pipeline to generate the synthetic travel demand from raw data, which can be extended and adapted by interested researchers or practitioners;
- synthetic travel demand, readily prepared for agent-based transport simulation, for the case of the Île-de-France region around Paris;
- thorough documentation of the synthesis process with basic data-driven approaches without the need for much calibration;
- the basis to ensure reproducibility of scientific agent-based studies in the Île-de-France region.

By that, we intend to foster open and reproducible research with agent-based models.

[1] Further models can be found at https://www.matsim.org/open-scenario-data

In addition, we invite researchers to test their specific methods for population synthesis, location assignment, and other aspects inside of an integrated pipeline. All code is available online (Appendix B), and the process is based on publicly available data. Methodologically, we intend to give ideas and first approaches for validating and verifying the quality of agent-based transport simulations, by examining error properties of the whole synthesis process.

The remainder of the paper is structured as follows. Section 3 describes the synthesis process. In particular, the data sets used will be covered in detail, as well as the methods used. Section 4 analyzes the generated population for Île-de-France in terms of fit to reference data. Afterward, Section 5 provides a more detailed analysis of the influence of sampling rate and confidence in the results. We finish with a discussion of the presented analyses in Section 6 and provide a final summary in Section 7. At this point, we would already like to point the reader to the glossary in Appendix A, which covers the abbreviations of data sets, concepts, and methods introduced throughout the paper.

3. Travel demand synthesis

Our synthesis pipeline aims to start with raw data sets, to transform them, to apply further models, and arrive at a final synthetic travel demand on a person level that can be readily used in a downstream agent-based transport simulation. We intentionally apply rather simple, data-driven algorithms to establish a baseline into which more sophisticated models can be integrated later on, with the ability to compare them against an established benchmark.

The proposed pipeline for Île-de-France consists of several steps that each draw from a specific data set and apply a particular algorithm to make use of the data in the synthetic travel demand. Figure 1 shows the general setup of this pipeline. In a sequence of steps, a data set is created that contains a representation of all households in the region, with individual persons attached to them. Those persons have a sequence of trips and activities, which they perform during an average workday in Île-de-France. While the household and person data sets contain rich sociodemographic information and their respective home locations, the activity data set contains the purposes, times, durations, and locations of those activities. The trips data set additionally adds the duration of trips and the chosen mode of transport.

For this synthesis process, data sets are required, which are also summarized in Figure 1. Except for the regional household travel survey, all data sets are publicly available and can be obtained by any researcher from the respective websites. As we also provide the pipeline code, as open-source software, it is, therefore, possible for anybody to recover the synthetic travel demand from raw data. While the quality can be improved, using the more fine-grained regional household travel survey, which is only available on request from the respective authorities, the national data set poses a viable open replacement.

Technically, the code is provided in two parts. The first part is `synpp`[2], a generic Python package for chaining algorithms and code pieces (*stages*) in a larger pipeline setup. While it can be used in a general way, it aims at providing a solid basis for travel demand synthesis and transport simulation applications.

The second part is the specific implementation of the Île-de-France use case, including the code for all data transformation, processing, and writing steps. The code is provided open source[3]. The Île-de-France pipeline is designed in a way that any researcher can download the necessary data sets and regenerate the synthetic travel demand for Île-de-France as it is described in this paper. Appendix B gives a first overview and directions on how to set up and run the pipeline.

In the following, Section 3.1 gives an overview of the data sources used, while Section 3.2 describes the methods which are applied to them.

3.1. Data sources and cleaning

In the following sections, data sets shall be presented, which make it possible to create a full synthetic travel demand for the Île-de-France region in France. The proposed process of synthesizing a travel demand, as outlined further below, is solely based on these data sets.

3.1.1. Spatial zoning system

In France, different spatial zoning systems are in use for statistical purposes. In this research, only a couple of them is used, mainly the ones that have the highest availability among the published data sets. The reference data set is a shapefile containing the contours of all IRIS zones in France. Those

[2] `https://github.com/eqasim-org/synpp`
[3] `https://github.com/eqasim-org/ile-de-france`

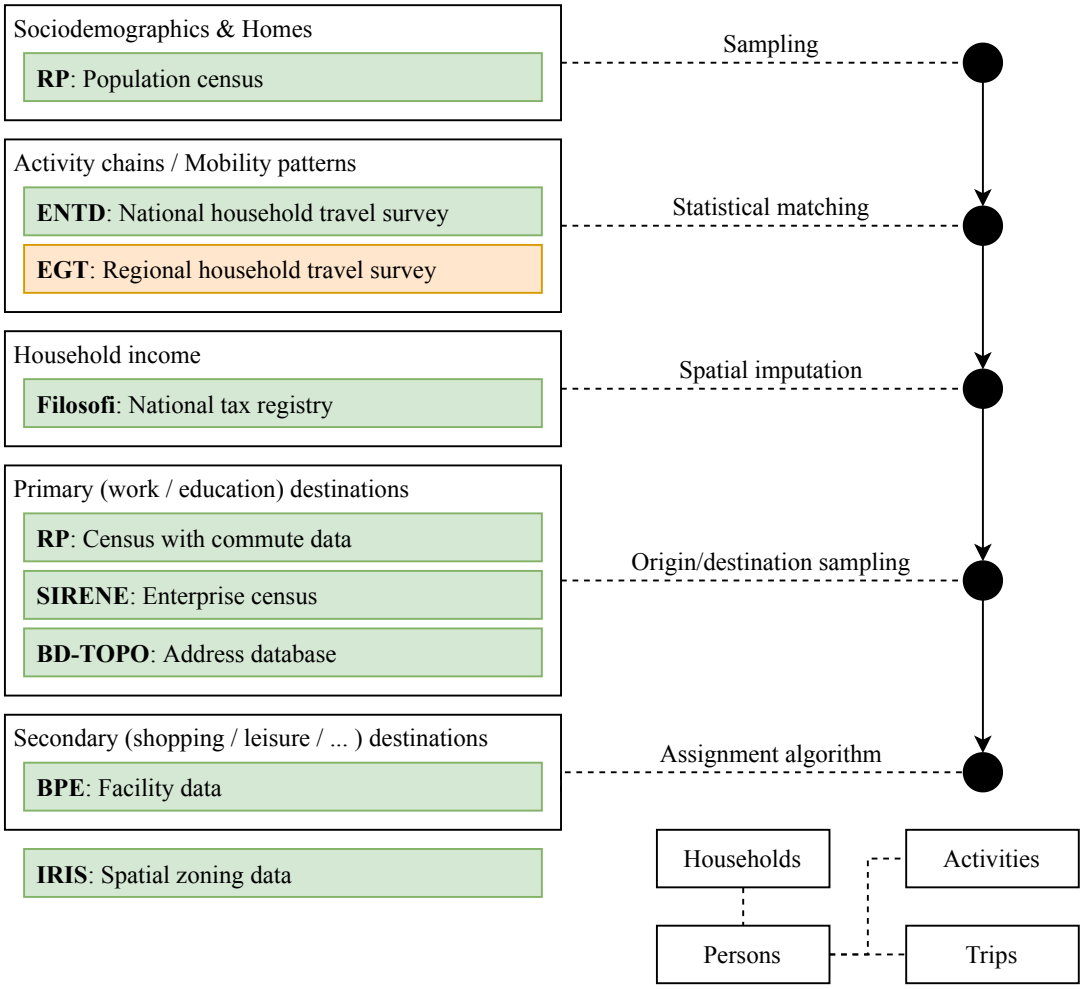


Figure 1: General setup of the synthesis pipeline with the used data sets. All data sets are open and publicly available (green), except the regional household travel survey (yellow).

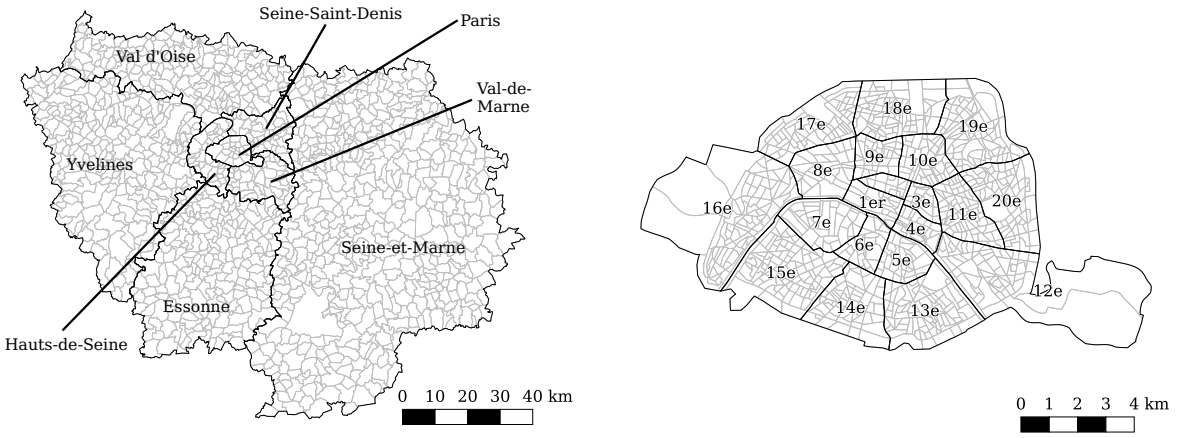
IRIS zones were introduced in 1999 and updated in 2008 to perform country-wide statistical analyses based on the national census. Each IRIS zone has an identifier that is divided into three parts: The first two digits denote the *département*, which is an upper-level administrational unit, the following three digits identify a *commune* (municipality) within this *département*, and the last four digits describe the statistical IRIS zone in that municipality. It is important to note that not all *communes* may be divided into IRIS, mainly if they have less than 10,000 inhabitants.

For the present study, only zones within the Île-de-France *region* in France are considered. Those are all which lie in the departments of Paris (75), Seine-et-Marne (77), Yvelines (78), Essonne (91), Hauts-de-Seine (92), Seine-Saint-Denis (93), Val-de-Marne (94), and Val-d'Oise (95). Figure 2a shows the area of the *départements* of Île-de-France and the *communes* into which they are divided. The city area of Paris is not divided into *communes*, but into 20 *arrondissements*. Their spatial extents are similar to those of the municipalities. Therefore, both *communes* and *arrondissements* are treated equally in the scope of our method. The segmentation of the Paris department into *arrondissements* is shown in Figure 2b. In this case, also the further division into statistical IRIS zones is shown.

For the Île-de-France region, we work with eight departments, 1,296 municipalities, and 5,259 IRIS covering an area of around $12,000$ km^2.

3.1.2. National census

National census data for France (*Recensement de la population*, RP) is published by *INSEE* (National Institute of Statistics and Economic Studies) on an annual basis three years after that data has been obtained. The latest available data set has been published in June 2018 and contains statistical information of a representative sample of the French population for the year 2015. For each household and each



(a) The *départements* of the study area and the *communes* into which they are divided.

(b) The *communes* (*arrondissements*) of the study area and the *IRIS* into which they are divided. Note that also the surrounding municipalities are covered by similarly sized IRIS.

Figure 2: Spatial zoning of the study area

Structural attributes	Person attributes	Household attributes	Spatial attributes
Household ID	Age	Household size	IRIS ID*
Person ID	Sex	Number of cars	Municipality ID*
Household weight	Employed (yes/no)		Department ID
	Ongoing education (yes/no)		
	Socio-professional category		

Table 1: Attributes per person resulting from cleaning and analysis of the French census dataset. *Spatial attributes are only given if available.

person, numerous sociodemographic attributes are given, such as age, gender, household income, number of cars, and others. A statistical weight is assigned to each household, which makes the data compatible with previous publications of the data set and other surveys performed in France.

For most households, the identifier of their home IRIS is given, which allows for the protection of person-specific data, but makes it also possible to use the data for synthesis purposes, as in this study. If an IRIS has less than 200 inhabitants, only the identifier of the municipality is given (0.07% of weighted households). Also, some municipalities are not covered by IRIS at all, because the municipality itself has a low number of inhabitants. In those cases, only the identifier of the department is known (10.06% of weighted households).

Nevertheless, the national census allows us to synthesize a population with spatially detailed sociodemographic attributes. It is also fortunate that the census is given on the household-level, with specific persons being directly attached to those households. This structure makes it possible also to synthesize realistic household-level distributions of sociodemographic attributes.

The census data set for 2015 features a large sample of individually weighted person observations for France and its overseas territories. For the Île-de-France region, it lists around 4.3 million residents in 1.9 million households, which makes around 35% of the real population.

For its use in our approach, the individual dataset is further cleaned. The resulting attributes can be seen in Table 1. Households containing persons who have their principal place of employment or education outside of Île-de-France are filtered out to restrict the study area to the region itself. These are 1.97% of the households containing 2.53% of all persons.

One attribute that requires further explanation is the *socio-professional category*. It is a well-defined (INSEE, 2003) concept which classifies persons into eight categories: (1) agricultural workers, (2) artisans, merchants, self-employed, (3) leading positions and intellectual workers, (4) intermediate professions, (5) employees, (6) workers, (7) retired, (8) others without employment.

In addition to the detailed data sets, INSEE also prepares aggregate datasets. These tables contain, for instance, an aggregated age distribution for each IRIS in France, including those whose inhabitants cannot be geolocalised in the individual data set. We use this dataset to enrich the zoning data with

information on the total number of inhabitants in each zone. Furthermore, we use the data as a reference when comparing the characteristics of our synthetic population.

3.1.3. Origin-Destination commute flows

Along with the census data, information about the commuting behavior of the French population is published by INSEE on an annual basis. For most of the individuals in the census data set, the origin-destination (OD) pair for the daily commute is recorded.

The information is provided in two data sets: One for work commute and one for commuting to educational activities such as school or university. In each case, the origin *commune* is given, along with the destination *commune*. Furthermore, the commutes are annotated with a statistical weight and the used mode of transport. It is not possible to reconstruct the actual individuals of the census data from these commutes.

As can be seen from Figure 2, commuting flows between municipalities can be understood as quite detailed from the perspective of the whole Île-de-France region. In Paris, however, this refers to the *arrondissements*, which renders the data set quite sparse. The potential of the data set is, therefore, not to produce highly realistic commuting patterns on a lower level like within the center of Paris. Instead, we use it to model the overall movement of people in and out of the city.

For the municipalities within Île-de-France, the data set contains around 8.3 million observations for work commutes and around 43,000 observations for education commutes.

Two OD matrices are derived from the data set, one for commutes to work and one for commutes to education. The process is straightforward: For each origin municipality, the weighted number of trips to each other municipality (destination) is tracked and divided by the total number of originating trips. This way, a probability of commuting to a particular destination municipality is established for each origin municipality. In only four cases (one for work and three for education), no trips are recorded at all for a specific origin. In these few cases, only trips inside of the same zone are allowed.

3.1.4. Income distribution

The Filosofi (*Fichier Localisé Social et Fiscal*) data set collects income data of tax registered people in France. It is published as open data three years after the acquisition of the data. The most recent data set has been published in 2018 and therefore contains income information of the population in 2015. Specifically, the data set provides the centiles of the distribution of *declared income* and *disposable income*. While the former mainly considers gross salaries, the latter takes into account deductions due to taxes, social security, state insurances, as well as social benefits. The distributions are given on the level of regions and municipalities, but one year after (i.e., four years after acquiring the data), a more fine-grained data set is published that provides distributions on the level of IRIS. The latter data set, however, does not provide further sociodemographic information, while the municipality-based version also offers income distributions by household size and a few other household-level sociodemographic attributes.

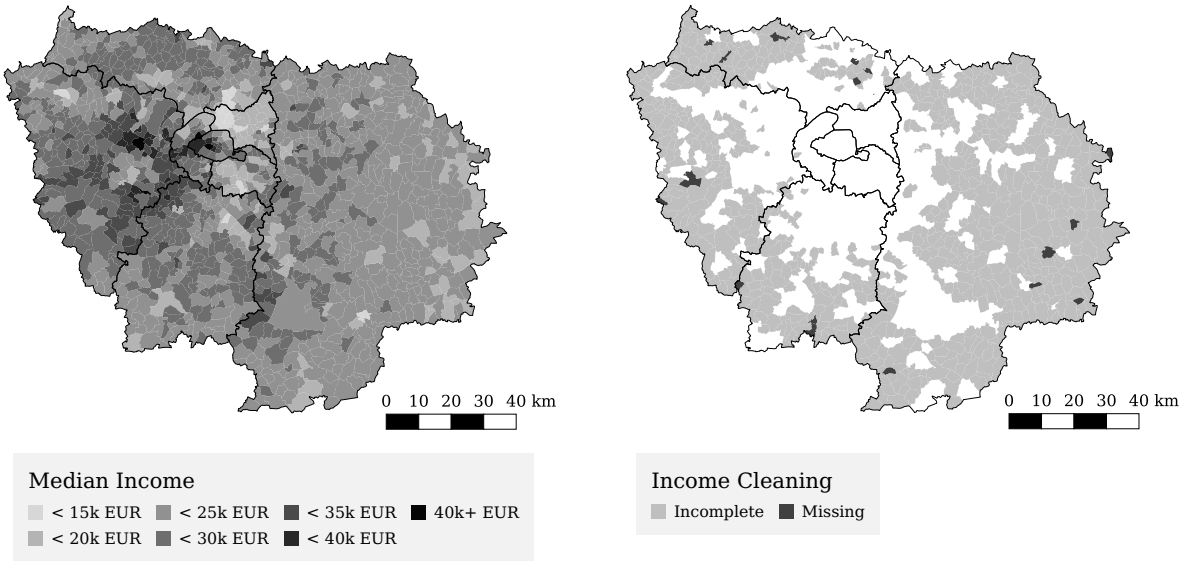
In the synthesis process, we use the regional Filosofi data set for validation. For synthesis, we only make use of the general income distributions by municipality. Further sociodemographic information could be included in the future.

The income provided in Filosofi is an annual income *per consumption unit*. It is a metric that makes incomes comparable between households. Since different household configurations entail different consumption patterns, the total income is not directly divided by the number of household members, but each of them is assigned a specific weight. INSEE, therefore, defines a consumption unit (*unité de consommation*) such that the first person over 14 years is counted as one full unit. Then, every additional person over 14 years is weighted by 0.5, while every person under 14 is weighted as 0.3. These definitions are important to arrive at meaningful comparisons of income distributions of different data sets. Income classes, for instance, in the household travel surveys (see below), usually refer to the *monthly disposable income per consumption unit*.

The distributions are given in eight centiles from the 10% to the 90% centile. Unfortunately, not all municipalities (715 of 1,296) in Île-de-France provide all centiles, mainly due to data protection for those areas with small population density. For these cases, only the median income in the municipality is known. Furthermore, 19 municipalities are not contained at all in the tax data set.

To clean the data, we first impute an income distribution to all municipalities for which only the median is known. We do so by comparing the median income values of the incomplete municipalities with the median income values of all known distributions. We then select the known municipality with the closest median and attach the full distribution to the incomplete municipality. In the future, more detailed imputation procedures could be applied, e.g., an additional matching by the Gini coefficient. Second, we

fix municipalities that are missing altogether by finding the nearest (centroid) neighbor municipality of each of them. We then attach the income distribution of the neighbor municipality to the missing one.



(a) Spatial income distribution across Île-de-France given as the *median monthly household income per consumption unit.*

(b) Visualization of missing information on municipal income distributions in Île-de-France. Incomplete municipalities only provide information on the median income, while missing municipalities provide no information at all.

Figure 3: Analysis of household income distributions in Île-de-France

Figure 3a shows the spatial distribution of median income per municipality. The median household income varies between around 13,000 EUR and 43,000 EUR across all municipalities. The overall median in Île-de-France (derived from the regional data set) is around 23,000 EUR. From Figure 3a one can see how household incomes are relatively higher in the west of Paris while incomes in the eastern suburbs are substantially lower. Figure 3b shows all municipalities for which some cleaning procedure was necessary. One can see that completely missing data is very rare, while municipalities with incomplete income distributions are located relatively far away from the city center of Paris. Especially, Paris itself and the three surrounding *départements* are covered well with detailed income distributions. The overall income distribution of Île-de-France is shown further below in Figure 6.

3.1.5. Household travel surveys

The *Enquête globale de transport* (EGT, Île-de-France Mobilités et al., 2010) is a household travel survey (HTS) conducted in the Île-de-France region, mainly during the year 2010. The survey has the classical structure of a household travel survey: Each member of a household is asked about their activities and travels during one particular reference day. Such surveys make it possible to estimate models of daily travel patterns, including the type of activities, the mode of transport for their connecting trips, and more.

The data set is divided into several parts that are relevant for the study at hand: For each household and each person, detailed sociodemographic information is available such as age and gender. Income classes by household are available. In a second table, one particular day is described for each of the persons by a chain of trips. Each trip holds information about the preceding and following activity, the mode of transport, distance, and duration. It must be noted that only activity chains for persons over five years old are recorded, while sociodemographic information is available for all. In terms of spatial information, the version of EGT that is available to the authors is relatively coarse since the trip start, and end locations are only given on the level of municipalities. In the best case, EGT could, therefore, be used to estimate OD flows between those zones, but for those, the census data set provides larger evidence. For the study at hand, the data set is a rich source of information because it defines the daily patterns of individual travelers. Since sociodemographic attributes are given, a connection to the census data can be established. Given a set of artificial agents with attributes such as age and gender, it is, therefore, possible to find activity chains with similar sociodemographics and attach them to those agents.

Furthermore, EGT gives a rich set of reference distributions, such as departure and arrival times during one day by mode of transport, mode shares in general or by the time of day, distances covered, and more. It, therefore, provides information that can help to validate the behavior of the synthetic population.

The EGT contains the trip chains of around 35,000 respondents in 15,000 households in the Île-de-France region. These numbers translate to a sample of around 0.3% of people living in the region. Within Île-de-France, around 122,000 trips are reported of all the members in each household.

Unfortunately, EGT is only available on request from the regional authorities and therefore not publicly available. As a publicly available alternative, the national household travel survey *Enquête nationale transports et déplacements* (ENTD) is available. It has been conducted between April 2007 and April 2008 and is, therefore, a bit older than EGT. Both data sets follow the same general structure, although available attributes and encodings vary slightly.

The big drawback of ENTD is that only 20,200 households in France were interviewed, which leaves 5,823 households with 14,216 persons for the Île-de-France region. Contrary to EGT, only one person per household is surveyed about their daily mobility pattern. We arrive at 4,613 respondents, which resembles only 0.04% of the population of Île-de-France. Therefore, the data set is much sparser than the regional travel survey. Persons under five years old are recorded in the households, but no specific activity chains are available.

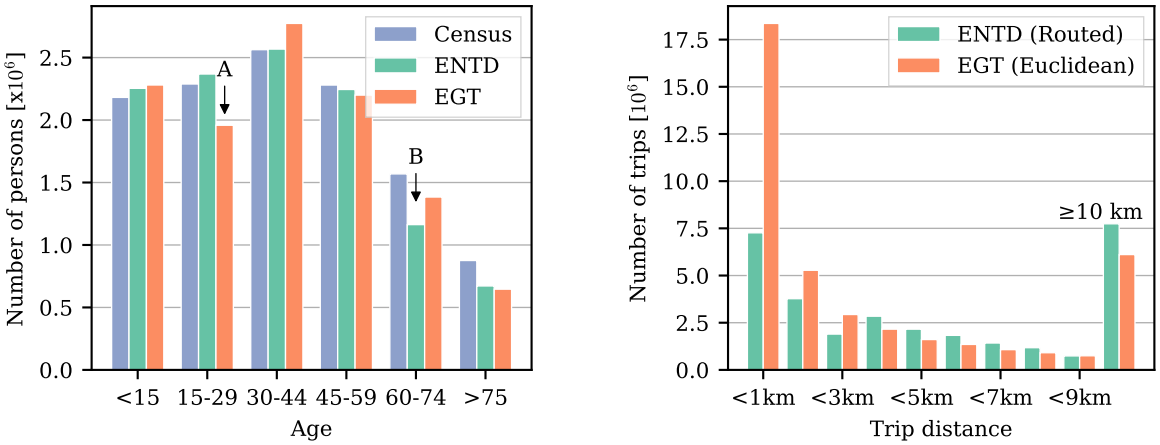
While most sociodemographic attributes of both ENTD and EGT fit well to the reference values of the census data, it must be pointed out that some differences exist. One interesting example is a shift in the age distribution, as is shown in Figure 4a. At point "A", a substantial number of 20-year-old people are missing in the regional EGT, while a large number of 60-year-old people are missing in the national ENTD. Explanations may be a different definition of "place of residence" in the data sets or differences in the weighting procedure or the chosen strata. Smaller differences (like for the age group of 33-40 years) may be explained by the way we filter for Île-de-France residents (see below).

ENTD and EGT both provide income classes for all households in the data set. While they are defined in different strata, they are also different in meaning. While ENTD provides household income classes per consumption unit, EGT provides classes of total household income. Since it is possible to calculate the consumption units in both data sets, it is possible to convert resulting income values into each other. Figure 6 shows a comparison between EGT, ENTD, and tax data by taking into account the midpoint of the respective income strata as the income value for each household. One can see the plain strata in ENTD as steps, while EGT is more smooth because the displayed income values are a result of dividing the midpoint value by the consumption units of each household. From Figure 6 we can see that there is (ignoring stratification) a strong resemblance between all three data sets.

On the trip level, ENTD yields an average number of 3.4 trips per day per active person (i.e., not staying at home) in Île-de-France. On the contrary, the regional EGT gives an average of 3.8 trips per day. For the department of Paris, the difference is even larger, with values of 3.4 in ENTD and 4.15 in EGT. Hence, there is a substantial difference that is difficult to explain only by a difference in time or sampling rate. Both numbers have been published previously for their respective data sets (INSEE, 2010; STIF et al., 2013). Figure 4b shows the distance distributions of both data sets. From the comparison, one can see that shorter distances are much less frequent in ENTD. This phenomenon is in line with the observation that for the same amount of (weighted) respondents, ENTD features only around three quarters, the number of trips that are reported in EGT. We, therefore, conclude that short trips are considerably underreported in ENTD, which makes EGT a more reliable source of trip-level information.

For the future, it might be interesting to understand further how the respective data sets have been weighted and from where those deviations may originate. However, for the work at hand, it is sufficient to have a set of activity chains, which are annotated with sociodemographic information. In that case, EGT will yield more realistic activity chains, because more of the short trips are reported. In the following, the ENTD data set is, therefore, used as an open and publicly accessible an alternative for the case where access to the EGT data is not possible or necessary. This may especially be the case if researchers only want to get started with the modeling system and only use publicly available data. Considering the distribution in Figure 4b, it may also be sufficient to use ENTD when rather large-scale use cases are in place, where trips under a range of around one to two kilometers are not expected to be affected substantially.

Inside of the synthesis pipeline, both data sets are cleaned such that they result in two structurally identical data frames. Table 2 summarizes the extracted attributes. Those attributes written in *italic* have the same set of possible values as the census data described above and can, therefore, be used for matching trips and activity chains to artificial persons. One exception is the *income class*, which is defined by different income strata in EGT and ENTD. However, this does not impose any restrictions



(a) Comparison of age distribution in the national household travel survey (ENTD) and the regional household travel survey (EGT).

(b) Comparison of trip distance distribution in the national household travel survey (ENTD) and the regional household travel survey (EGT). Note that ENTD is shown with routed trip distances, while EGT is shown with Euclidean distances. They are compared as no data set provides information on both distance metrics.

Figure 4: Comparison of ENTD and EGT

Structural attributes	Person attributes	Household attributes
Chain ID	*Age*	*Département ID*
	Sex	*Household size*
	Employed (yes/no)	*Number of cars*
	Ongoing education (yes/no)	Consumption units
	Socioprofessional category	Income class*
	Driving license (yes/no)	Number of bicycles
	Public transport subscription (yes/no)	

Table 2: Attributes per person resulting from cleaning and analysis of the French HTS datasets. Attributes in *italic* have the same structure as the census attributes in Table 1. *Income classes are defined as different income strata in EGT and ENTD.

on the matching process, as will be shown further below. Additionally, the structure of both HTS allows calculating the consumption units per household as defined previously.

On the trip level, both HTS are cleaned to provide the following attributes: departure time, arrival time, the purpose of the following activity, the purpose of the preceding activity, mode of transport, origin department, destination department, and distance. Note that EGT only provides Euclidean distances, while ENTD only provides routed distances. Figure 5 summarizes the available information plus additional information such as trip and activity durations, which can be easily derived from the data. We consider *car driver*, *car passenger*, *public transport*, *bicycle*, and *walk* as modes of transport. While both HTS allow for a more detailed analysis, those are the ones that we consider essential for a first version of the model that can be refined later on. For activities, we use the specific types of *home*, *work*, *education* (which will be referred to as *primary activities* in the following), and *leisure*, *shopping*, and *other* (*secondary activities*). The same concept applies as for the modes of transport: A much more fine-grained distinction would be possible later on, but is not considered in the basic model. For instance, introducing a distinct *food* category for trips to a restaurant or lunch break could be a straightforward extension.

Finally, both HTS are cleaned to reference the same group of people only staying in Île-de-France: Persons with trips that go beyond the border of Île-de-France are deleted from the data set, as well as those which do not have their residence in the area (only applicable to ENTD).

At this point, it should be noted that currently (April 2020), new versions of both EGT and ENTD are under preparation, with the new ENTD performed between April 2018 and April 2019 expected to be published by 2021.

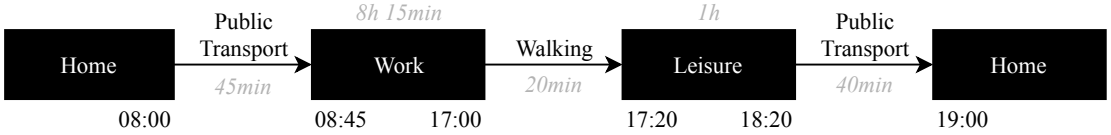


Figure 5: Example of an activity chain with available attributes from French HTS data. Derived attributes are shown in light gray.

3.1.6. Address database

The National geographic institute of France (IGN) provides a regularly updated and publicly accessible database of all registered addresses in France called BD-TOPO. It contains the written address including street name and house number, as well as a distinct coordinate for each observation. For Île-de-France, the data set contains 2,131,728 individual addresses, of which 1,891,175 can be used in our process because they have valid street names, house numbers and municipality identifiers. The data set allows us to define locations of daily activities in a detailed way.

3.1.7. Enterprise census

In France an open an publicly accessible enterprise census exists (SIRENE, *Système national d'identification et du répertoire des entreprises et de leurs établissements*). The SIRENE data sets lists all enterprises registered in France to date and in the past, updated every month. It further divides enterprises into individual facilities with unique identifiers. For each enterprise and facility it provides the number of employees and the type of sector according to the official enterprise classification system of France (NAF, *Nomenclature d'activités française*). While the data set provides the address of each facility in written form, as well as the identifier of their municipality, their location is not known by coordinate in an easily digitally processable way.

Therefore, we match the 411,608 available facilities in Île-de-France with the address database. 379,175 of addresses can be matched exactly to the coordinate by using street name, house number and municipality identifier. 7,521 can be matched without taking the municipality into account (to fix cases in which the address database and the enterprise database may be out of synch due to municipality mergers or separations). Finally, an additional 5,467 are matched by municipality identifier and a Levenshtein distance (Levenshtein, 1966) compared to a candidate from the address database of at most five modifications. In total, that gives 392,163 (95.28%) enterprises that can be geolocalized on three different levels of confidence.

The enterprise census data is used to define work places for the synthetic population in detail.

3.1.8. Service and facility data

A service and facility census (BPE, *Base permante des équipements*) is published on an annual basis by INSEE. It consolidates several independent data sets with the goal of establishing a central registry that lists services and facilities with their location and type in France. While many are annotated with exact coordinates, some are only known by IRIS or municipality.

During the clean-up process, all observations are deleted, which do not provide either a valid IRIS or municipality identifier within Île-de-France. All remaining observations that do not provide exact coordinates are placed at a random address (see above) inside of their respective IRIS if it is specified, otherwise inside of their associated municipality. After applying this process, we arrive at 469,181 facilities.

The cleaned *BPE* allows us to assign the location of agent activities such as shopping realistically during demand synthesis. Analogously to the activity types mentioned above, we divide the facility type into four categories: *education* (11,267 obs.), *shop* (67,458 obs.), *leisure* (64,416 obs.), and *other* (326,040 obs.). Note that later on, all of these facilities will be treated as *work* locations. In theory, also the *BPE* would allow for a much more fine-grained definition of activity types, which offers the potential to improve the synthesis process in the future. The category *other* currently mainly consists of facilities from the sectors of health, transport, and tourism.

3.2. Synthesis process

The data sets described above allow us to synthesize an artificial population of travelers. The main component of this process is a statistical matching procedure that combines data from the census data with the selected household travel survey. For that purpose, the census data is furthermore spatially

enriched with the income distribution data. Finally, the synthesized persons need to be assigned locations for their primary and secondary activities. Those steps are detailed in the following sections. While we present straight-forward algorithms to establish a baseline for future developments and comparisons, we give pointers to more advanced methods which could be integrated later on.

3.2.1. Population sampling

Population synthesis is the process of generating a set of households and persons with sociodemographic attributes. Commonly, the major task of a population synthesis algorithm is to process a small sample of the population (for instance, from a household travel survey) to create a model from which the full population can be generated under certain assumptions. The most common algorithm is Iterative Proportional Fitting (IPF), where each individual in the population sample is assigned a weight such that the weighted population shows predefined marginal distributions for age, gender, and other attributes or combinations thereof (e.g. Arentze et al., 2007; Rich and Mulalic, 2012). As an extension, Iterative Proportional Updating (e.g. Pendyala et al., 2012) weighs households to match person and household level attributes. An overview of such fitting methods gives Müller (2017). Methods based on Monte Carlo simulation have been proposed by Farooq et al. (2013) and adapted by Saadi et al. (2016b) presenting a Hidden Markov Model where persons and attributes are sampled one after another and dependent on previous states. Saadi et al. (2018a) provide a comparison of fitting-based and sampling-based approaches. Hierarchical models for the sampling of person-level, household-level and household member-level attributes are proposed by Saadi et al. (2018b) and Sun et al. (2018). A related line of research is the use of Bayesian Networks to graphically represent and leverage the interdependency between person and household attributes (Sun and Erath, 2015). Recently, approaches making use of Deep Generative Modeling (DGM) have been proposed (e.g. Borysov et al., 2019; Garrido et al., 2020). While these methods are mainly necessary to scale up and enrich relatively sparse, small census samples or household travel surveys, the data available for France is suited for direct sampling, as will be shown below. Yet, including any of the above-mentioned methods could give modelers the power to manually design future scenarios by proposing new marginal distributions of attributes (e.g. in IPF) or by changing relationships between attributes (e.g. in a Bayesian Network).

Since the census data described in the previous section is only a sample of the whole population of Île-de-France, we need to scale it up to arrive at a full set of agents. Fortunately, the census (household) weights $w_i \in \mathbb{R}^+$ provided by INSEE make this procedure easy. In theory, each of the existing households i represents w_i households in reality, which means that we can copy them w_i times. As those weights are not integers, we apply stochastic rounding (Gupta et al., 2015)[4] to arrive at integer multiplicators $m_i \in \mathbb{N}^+$ for each household:

$$m_i = \begin{cases} \lfloor w_i \rfloor & \text{with probability } 1 - (w_i - \lfloor w_i \rfloor) \\ \lfloor w_i \rfloor + 1 & \text{with probability } w_i - \lfloor w_i \rfloor \end{cases} \tag{1}$$

This process of obtaining household multiplicators is designed to be deterministic given a random seed R. After obtaining the multiplicators m_i, we arrive at a full population of Île-de-France that contains around 12 million persons with sociodemographic attributes.

In many cases, simulations can not be performed with such large populations due to runtime constraints. Therefore, this population is optionally scaled-down afterward. For that, we define the sampling rate s. Downsampling of the population follows a straightforward process. The algorithm goes through all households step by step and keeps each of them in the final population sample with a probability of s. A sampling rate of $s = \frac{1}{2}$ would, therefore, mean that a synthetic population of half the size of the real population is sampled, while a value of $s = \frac{1}{100}$ would yield a 1% sample.

Finally, each of the synthetic households and persons is assigned a new unique identifier within the synthetic population. For later analysis, the originating census identifiers are kept in the data set.

3.2.2. Home location assignment

As described initially, the census data contains "zeros" with regards to the home location of households. Therefore, this is also true for the sampled agent population. Nevertheless, we would like to assign a specific home coordinate to each of the artificial households.

[4] In the context of population synthesis and Iterative Proportional Fitting, the method has also been labeled as *Truncate, Replicate, Sample* by Lovelace and Ballas (2013).

In the data, we observe three cases: First, there are cases in which neither municipality nor IRIS of the household is known. These are those municipalities that have not been divided into IRIS by the statistical office due to low population density. Second, some households have information about their municipality, but not about the IRIS. These cases represent municipalities that are rich in population in general, but the respective IRIS has a population of fewer than 200 inhabitants. Finally, we have households for which municipality and IRIS are known. These are the cases that do not need further corrections.

The data of the two other cases are augmented as follows. Since, in any case, we have information about the *departement* of the household, we can first select all municipalities in a household's *departement* which are not covered by IRIS. We then weigh them by population density, which comes from the aggregated census data provided by INSEE. We then draw one municipality from this distribution. For the second case, we follow a similar procedure. Here, we know the municipality of the household, so we can select all IRIS that have less than 200 inhabitants within this municipality. Again, they can be weighted as we know the exact population count from the aggregated data set. We then draw one IRIS from this distribution and repeat the procedure for all households.

Finally, we arrive at a population in which every household is assigned to a well-defined zone. A random address with coordinate for the address database is sampled for each household in its respective area to complete the process. Thus, given the zone information and the address database, the process of assigning a home location is rather easy. In case these attributes were not given, additional data, such as land-use data or satellity imagery could be used (Chapuis et al., 2018).

3.2.3. Income assignment

So far, the synthesis pipeline only considers income distributions by municipality. The process of attaching incomes to households is, therefore, relatively simple. For each household, the residence municipality is already known from the home location assignment step. This information is then used to find the respective municipal income distribution for each observation. Those distributions are given in deciles. First, one of those deciles is sampled for each household in a municipality. Afterward, a random income value is sampled from the range between the lower and upper bound of each household's decile. Note that the attached income values are *household incomes per consumption unit*, which means that household size is implicitly taken into account in this procedure. As the consumption units are known for each household, we can derive the total household income afterward. Figure 6 shows the resulting income distribution for the population with a good fit to the referential tax data set.

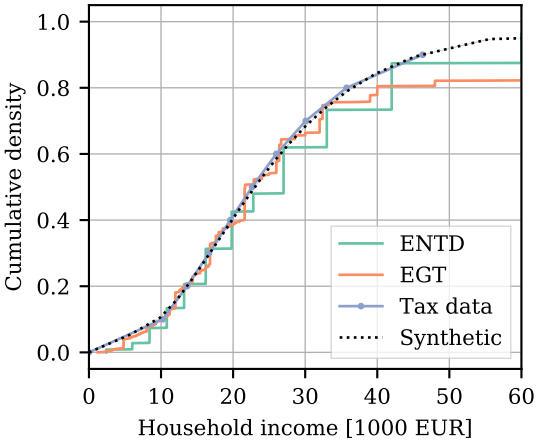


Figure 6: Comparison of income distributions in the national household travel survey (ENTD), the regional household travel survey (EGT), tax data (Filosofi), and the synthetic population. The income is given in *annual household income per consumption unit*.

3.2.4. Statistical matching

After income assignment, all data is in place to combine the sampled synthetic population with data from the selected HTS. The aim of this process is two-fold: First, some interesting attributes may not be available in the census and the synthetic population at this stage. For instance, it is not known which of the synthetic persons has a driver's license. However, this information is known for every

observation of the HTS. At the same time, both data sets feature attributes that have the same meaning and set of discrete values (such as age class, sex, income class). Assuming that those mutual attributes are sufficiently correlated with the unilateral attributes of interest, it is possible to enrich the synthetic population with additional information. This enrichment can be done by finding HTS source observations that are sufficiently similar to the synthetic target persons in terms of mutual attributes. Their unilateral attributes can then be attached to the target persons (and, in principle, households).

Second, each person from the HTS comes with a full day activity and trip chain. Assuming again that the sociodemographic attributes are sufficiently correlated with daily activity patterns, we can pose the same argument as above to attach whole chains from HTS data to the synthetic population.

In technical terms, we apply a procedure that is inspired by *statistical matching* algorithms (D'Orazio et al., 2006). First, we define a list of matching attributes $A_{1:N}$. For each (*target*) observation t of the synthetic population, it is then possible to note down a vector of attribute values:

$$\mathbf{a}_t = (a_{t,1}, ..., a_{t,N}) \tag{2}$$

Likewise, every HTS (*source*) observation can be identified by an index s with the respective attribute vectors:

$$\mathbf{a}_s = (a_{s,1}, ..., a_{s,N}) \tag{3}$$

Additionally, a weight w_s is known for each source observation.

In the following, we describe the matching algorithm for a given target observation t. The idea is to find all source observations that match in all predefined attributes and use their weights to sample one of them. However, this is an ideal case as the required combination of attributes may not even be available in the source sample. Furthermore, we seek to avoid overfitting by drawing from a very small set of source observations. Let $\mathcal{S}_t^k$ with $k \in \{1, ..., N\}$ define the set of source observations that match in the first k attributes to the target:

$$\mathcal{S}_t^k = \{s \mid \mathbf{a}_{s,1:k} = \mathbf{a}_{t,1:k}\} \tag{4}$$

We then define the actual selection set level k^* as the one that allows us to draw from at least M source observations:

$$k_t^* = \max\left\{k \;\middle|\; |\mathcal{S}_t^k| \geq M\right\} \tag{5}$$

Let $\mathcal{S}_t^* = \mathcal{S}_t^{k_t^*}$ be the final set of candidates for target observation t with which a probability density over the source sample can be constructed:

$$\pi_t(s) = \begin{cases} w_s / \sum_{s' \in \mathcal{S}_t^*} w_{s'} & \text{if } s \in \mathcal{S}_t^* \\ 0 & \text{else} \end{cases} \tag{6}$$

Using the density, one source observation s_t^* can be sampled, and the whole process is repeated for each target observation t. It should be pointed out that this algorithm can be heavily parallelized and optimized[5] on the implementational side, which allows for fast execution speeds.

For the specific case of Île-de-France, we choose the matching attributes *age class* (see Figure 4a), *sex*, and *socioprofessional category*. Those attributes get matched for 100% of the persons in a typical run of the algorithm. Additionally, we control for whether the household has *any cars* with a usual matching rate of 98%, *income class* (67%), and *departement* (18%). The minimum number of source observations is set to $M = 20$.

Once more, it should be noted that the proposed algorithm merely replicates the current mobility patterns of the population of Île-de-France. This is a common pattern which is also used in other research, e.g. by He et al. (2020) where activity chains are attached to persons based on their place of residence and occupation. More often, however, synthetic popualtions are passed to activity-based models (Arentze et al., 2007; Pendyala et al., 2012) where a sequence of statistical models is applied to construct activity chains step by step. Recently, more data-driven methods have been proposed, based on protein sequencing methods, namely Sequence Alignment Models (SAM, Shoval and Isaacson, 2007) and Profile Hidden Markov Models (pHMM, Liu et al., 2015; Saadi et al., 2016a). Using an approach with privacy

[5]Optimization is achieved using the powerful *numba* (*http://numba.pydata.org*) JIT compiler for Python as well as predetermining and sampling of all random numbers that are required in the sampling step upfront.

in mind, Ballis and Dimitriou (2020) propose a method to reconstruct activtiy chains from aggregated origin-destination matrices for different times of the day (which could be a result from anonymizing surveys or mobile phone data). Joubert and de Waal (2020) present a Bayesian Network approach and highlight the large potential of having a behavioural model that can easily be tuned by experts and planners.

3.2.5. Primary location assignment

As primary activity types, we consider *home*, *work*, and *education*. For those, the census data provides additional commute information in the form of commuter matrices. These matrices describe how many people would commute from a particular home municipality to all municipalities in Île-de-France. While the data allows separating the *work* commutes by mode, we do not consider this distinction for now.

The aim of the primary location assignment is twofold: First, the correct number of people should commute from one municipality to another in the population; second, the commute distance should fit the activity chains that have been assigned to each agent in the previous step.

The synthesis step is performed in three stages. In the first stage, we iterate through all municipalities and determine how many people in the synthesized population have their home located inside of each municipality k and need commuting information. Whether this is the case is determined by examining whether the person has a *work* or *education* activity, respectively, in their assigned activity chain. While the following process is executed for *work* and *education* commutes, we present it in general terms as the algorithm is applied independently for both activity types. The counting process results in a demand number O_k for each municipality. From the data cleaning part, we have already obtained a commuter matrix, which gives a probability $\pi_{k,k'}$. It describes the likelihood that a commute trip from origin municipality k to destination municipality k' exists. For each origin municipality k, we can therefore sample trip counts $f_{k,k'} \in \mathbb{N}^+$ to the destination municipalities from a multinomial distribution

$$(f_{k,1}, ..., f_{k,\cdot}) \sim \text{Multinomial}(O_k; \pi_{k,\cdot}) \tag{7}$$

such that in total we arrive at $\sum_{k'} f_{k,k'} = O_k$. Note that at this point, we have sampled an abstract mass of commute trips, which are not yet assigned to specific synthetic persons.

The second step is to find specific commute destinations for each of the sampled trips. For a combination of municipalities (k, k') we sample $f_{k,k'}$ random destination candidates with replacement among all enterprises available in the destination municipality. The number of employees is used as the sampling weight. We can then define the sampled set of candidates as $\mathcal{C}_{(k,k')} = \left\{c_{(k,k'),1}, ..., c_{(k,k'),f_{k,k'}}\right\}$.

Finally, destinations can be assigned to the synthetic population. To do so, a combined set of destination candidates is constructed for each home municipality k: $\mathcal{C}_k = \bigcup_{k'} \mathcal{C}_{k,k'}$. Note that the locations in this set are still spatially distributed in a way that they resemble the commute probabilities between the municipalities. For municipality k, we can now determine all persons with a home located in that municipality and refer to them via an index $u \in \{1, ..., O_k\}$. Likewise, we can refer to the candidates in $\mathcal{C}_k$ in an ordered way through $v \in \{1, ..., O_k\}$[6]. With the notation at hand, the assignment of a commute destination to a person can now be described by a *bijective* mapping $A : v \mapsto u$, i.e., for each person one destination must be chosen and each destination must be chosen exactly once.

The most simple mapping is $v = A(u) = u$, which corresponds to a random assignment of destinations for municipality k among all inhabitants. Unfortunately, this can lead to inconsistent situations as the assigned activity chains for the persons may contain trips to work, which are very short (in terms of travel time and distance from the HTS), but may still be assigned a commute destination far away. Therefore, we initially determine a *commute distance* for each person u by finding the first trip in their HTS-based activity chain that takes place between a *home* and a *work* (or *education*, respectively) activity. After that, we denote the distance as the commute distance $d_u \in \mathbb{R}^+$. In some cases, agents may have *work* in their activity chain, but not direct trips between home and work. As the HTS does not provide information about the actual locations of the activities, we assign commute distances randomly to those agents, by performing a weighted sampling among the commute distances that are known from the HTS according to the described approach.

At this point, the expected commute distance d_u is known for each agent, but also the home location $h_u \in \mathbb{R}^2$ from the previous synthesis step. Also, the location of each destination candidate $l_v \in \mathbb{R}^2$ is known. We then define the destination mapping A according to Algorithm 1.

[6]Note that $O_k = |\mathcal{C}_k|$

Algorithm 1 Primary location assignment

Initialize ordering $V = (-1, ..., -1)$ with $|V| = O_k$
for all persons $u \in \{1, ..., O_k\}$ **do**
  Initialize costs $J = (\infty, ..., \infty)$
  for all destination $v \in \{1, ..., O_k\} \setminus V$ **do**
    $J_v = \text{abs}\left(\|h_u - l_v\| - d_u\right)$
  end for
  $V_u = \arg\min_v J_v$
end for

This algorithm iterates through the persons one by one and calculates the distance between its home location and all available destinations. An offset is then calculated between the resulting commute distance and the desired commute distance d_u. In case a destination was already assigned to a previous person, the offset value is kept at infinity. The destination with the smallest offset in commute distance is then assigned to the agent. This procedure makes sure that agents with short commute distances in their plan are likely to be assigned a destination that is not too far away from the home location.

Finally, the assigned locations are saved with the synthetic persons for future analysis and further processing in the next synthesis step. Other approaches could be used for the assignment of primary activities. There is a long tradition of using gravity models (Jensen-Butler, 1972; Ahrens and Lyons, 2020) which quantify origin-destination flows by production and attraction of each zone. Those are quantified by relation-based attributes such as travel time and zone-based attributes such as land use or employment. Vitins et al. (2016) use a discrete destination choice model inside an optimization problem to consider capacity constraints of work places. Fournier et al. (2020) propose an integrated model of population synthesis and work place assignment using IPF and origin-destination-industry (ODI) matrices.

3.2.6. Secondary location assignment

In the last synthesis step, locations for secondary activities (in the current case *shop*, *leisure*, and *other*) are assigned using a specially designed algorithm. For a detailed analysis, we would like to point the reader to (Hörl and Axhausen, 2020) while here, only the basic idea is covered to give an intuition about the algorithm.

The activity chains of the synthetic population consist of information on the assigned trips and activities as well as the locations of primary activities. Each chain, therefore, represents a skeleton in which specific (i.e., primary) activities are fixed, and others are not assigned a location yet. However, from HTS data, it is known which travel time the trips between all activities should *ideally* have.

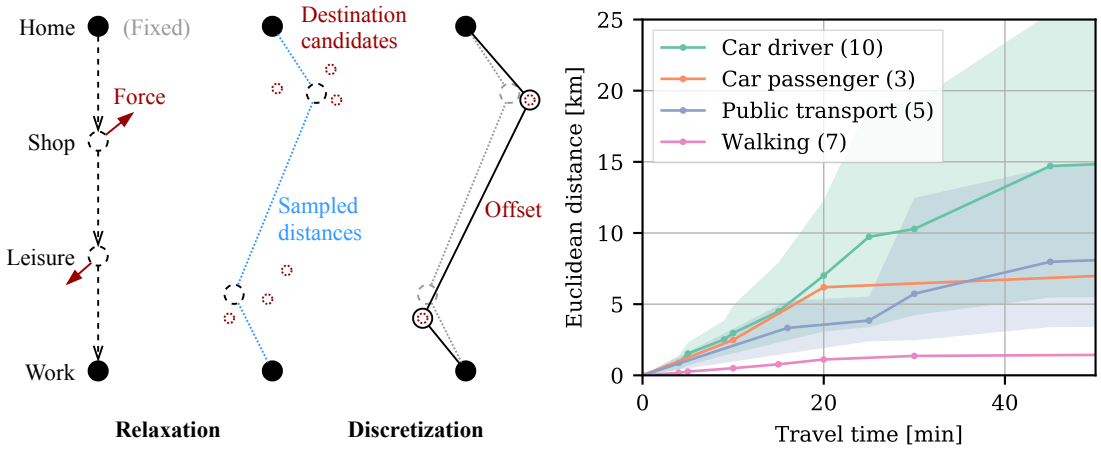
The algorithm first searches for all *assignment problems* in the activity chains, which are characterized by a fixed activity, followed by various assignable activities, followed by another fixed activity. In most cases, there is only one variable activity between two fixed ones, but more complex activity chains exist. Each assignment problem is treated separately.

As a preparation for the algorithm, all trips in the regional HTS are analyzed and divided into bins of modes and travel times such that each combined bin contains at least 200 observations. Figure 7b shows this analysis for the regional HTS: Given a particular mode and travel time, a distribution of distances is available.

In the first "relaxation" step of the assignment algorithm, the nominal travel times of each trip along the assignment problem are considered. Given the transport mode of each trip and the travel time, a distance is sampled from the prepared distance distributions. Afterward, a particle gravity model is used to place the variable activities in Euclidean space such that the distances between all coordinates resemble the sampled distances. In the second "discretization" step of the process, the closest destinations (from the service and facility data) are obtained, which are consistent with the activity types of the variable activities. For instance, a shopping activity will be snapped to the closest shop from its location in Euclidean space. The algorithm then follows a couple of convergence criteria, for instance, the offset of the resulting distances from the initially sampled ones and is called iteratively in case the criteria are not fulfilled. This process is summarized again in Figure 7a.

In essence, this algorithm assigns discrete locations to all remaining secondary activities while maintaining realistic distance distributions given the travel times and modes in the activity chains.

The secondary location assignment step is the only point in the pipeline were a limited calibration is taking place. In fact, the time/mode-bin-based distributions from HTS are not used directly in the algorithm, but small distances are oversampled. This is necessary as the algorithm tends to oversample



(a) Overview of the discretization-relaxation algorithm by Hörl and Axhausen (2020) used to assigned secondary activity locations in the Île-de-France population. First, activities are moved to locations that are consistent with a set of sampled distances; second, they are snapped to a priori know locations of discrete candidates. Dependent on the offset between the sampled distances and the resulting distances from the discretized locations, the algorithm terminates or is repeated.

(b) Input data to the secondary location assignment algorithm by Hörl and Axhausen (2020). For each mode, all trips from the regional HTS are divided into travel time bins such that each bin contains at least 200 observations. Inside of the brackets, the resulting number of bins is shown. The mean for the travel-time-binned distance distributions is shown. For "car driver" and "public transport", the shaded area shows the 90% confidence bounds of the respective travel-time-binned distributions.

Figure 7: Secondary location assignment

long distances because of structural constraints. To counteract, the distance weights f inside each time-bin/mode distribution are adjusted as

$$f'(d_i) = \begin{cases} f(d_i) \cdot (1 + \alpha \cdot (i/N)) & \text{if } \alpha \geq 0 \\ f(d_i) \cdot (1 + |\alpha| \cdot [1 - (i/N)]) & \text{else} \end{cases} \tag{8}$$

with the ordered distances $d_i \leq d_{i+1}$, the number of distances N in the respective time/mode bin, and calibration factors $\alpha \in \mathbb{R}$. For the population of Île-de-France, the values $\alpha_{car} = 0$, $\alpha_{\text{carPassenger}} = \frac{1}{10}$, $\alpha_{\text{pt}} = \frac{1}{2}$, $\alpha_{\text{bicycle}} = 0$, $\alpha_{\text{walk}} = -\frac{1}{2}$ are used.

We use the presented algorithm as it only poses very few restrictions on the available data. While more elaborate algorithms exist, they often require more complex calculations and almost always make use of a-priori known travel times between zones or locations in the study area to determine the attractiveness or likelihood of certain places. In our modeling set-up these travel times are the outcome of the downstream transport simulation after demand generation. A pathway for future development could therefore be to use more advanced models with initial best-guess travel times, run the population synthesis, obtain updated travel times, and to repeat this process iteratively. There is a range of methods that could be used, e.g. solving a Traveling Salesman Problem with Time Windows (TSP-TW) to find viable activity locations between two fixed ones (Esztergár-Kiss et al., 2018, 2020). Most approaches, however, make use of the space-time prism concept: In a first step, all locations are determined which allow to reach a discretionary activity from home, to perform the activity at this location (with a certain duration) and to continue the journey to the work place such that the person arrives on time and certain constraints on opening times or daily travel time budgets are fulfilled (Yoon et al., 2012; Justen et al., 2013). Afterwards, discrete choice models are applied to choose one candidate from the determined choice set. Recently, Ma and Klein (2018) have used Bayesian Network to devise certain heuristic choice strategies which are used instead, based on sociodemographic attributes of the travellers.

4. Validation

In the following an analysis of the generated travel demand based on the open national household travel survey (ENTD) is shown. While later on (Section 5) we present a more detailed analysis of the impact of different sampling rates, we stick to a value of 5% here. This sampling has been shown in preliminary experiments to provide acceptable run times of the downstream agent-based simulation while preserving the main statistical properties of the demand. Furthermore, at a sampling rate of 5% it is possible to show some stochastic variability in the outcomes. To account for this variability, we

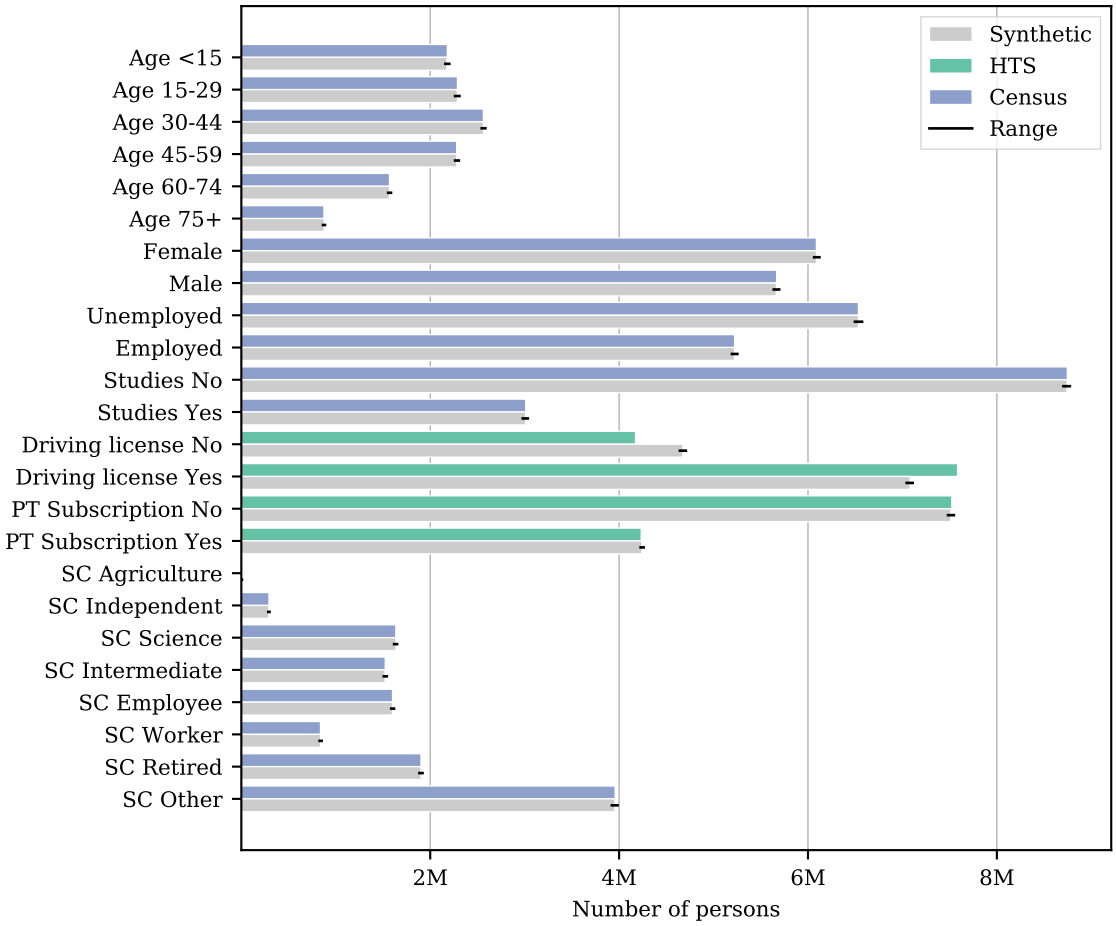


Figure 8: Comparison of person-level attributes between the synthetic population and reference data for a 5% sampling rate.

generate 200 synthetic travel demand configurations with random seeds $R = k \times 20$ with $k \in \{1, ..., 20\}$. The graphics below show the mean of certain quantities (e.g., the number of persons in a stratum) and the range between minimum and maximum inside the sample of size 200.

The sociodemographic structure of the population is mainly dependent on the population sampling and statistical matching steps in the synthesis process. Figure 8 shows a comparison between various population strata between the synthetic travel demand and reference data. One can see that the attributes that were generated from census data match well, even if the population is downsampled. Several attributes are not known from census data but attached through the matching process from the household travel survey. While some attributes fit well with the reference data (for instance, whether the persons have a public transport subscription), other attributes do not match precisely in the mean. This behavior is expected as there is no guarantee in our matching algorithm to arrive at consistent estimates. Hence, if statistically significant analyses should be performed on such attributes, more elaborate algorithms would need to be used. The same is true for household-level attributes, as shown in Figure 9.

Besides the overall validation of sociodemographic attributes, it is interesting to know whether the population spatially represents important features correctly. Figure 10 shows a comparison between the 13th arrondissement in the center of Paris with the suburban municipality Alfortville. From the distributions, one can see that the 13th arrondissement features more single households than Alfortville. Furthermore, the age distribution shows peaks for people between 15 to 29 years old with few under 15 years old, while Alfortville shows a larger number of very young and a large number of 30 to 44-year-old people. Hence, it becomes evident from the data that Alfortville is a more family-centric area. This fact is represented well in the synthetic population.

Also, Figure 10 shows that a sampling rate of 5% leads to substantial variance among different realizations of the population on a more local level. Increasing the sampling rate or averaging over multiple realizations would help to increase the confidence in the presented marginals.

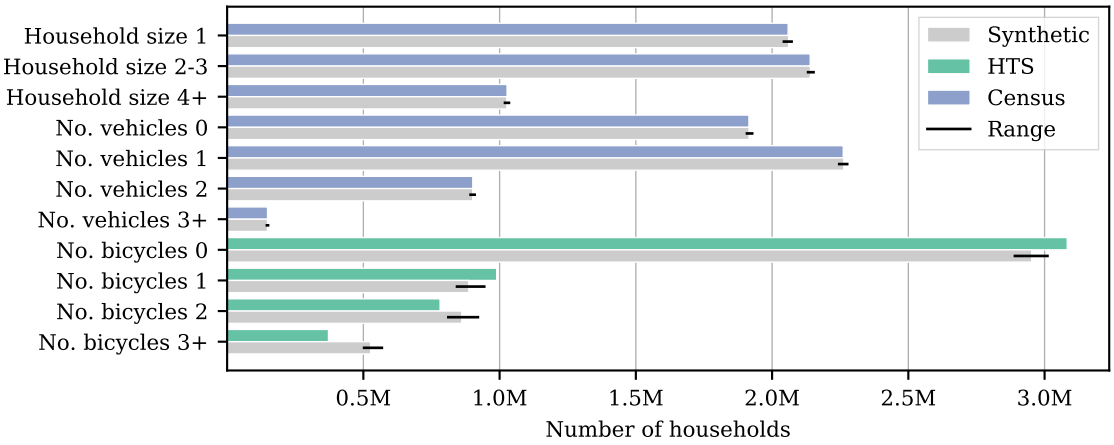


Figure 9: Comparison of household-level attributes between the synthetic population and reference data for a 5% sampling rate.

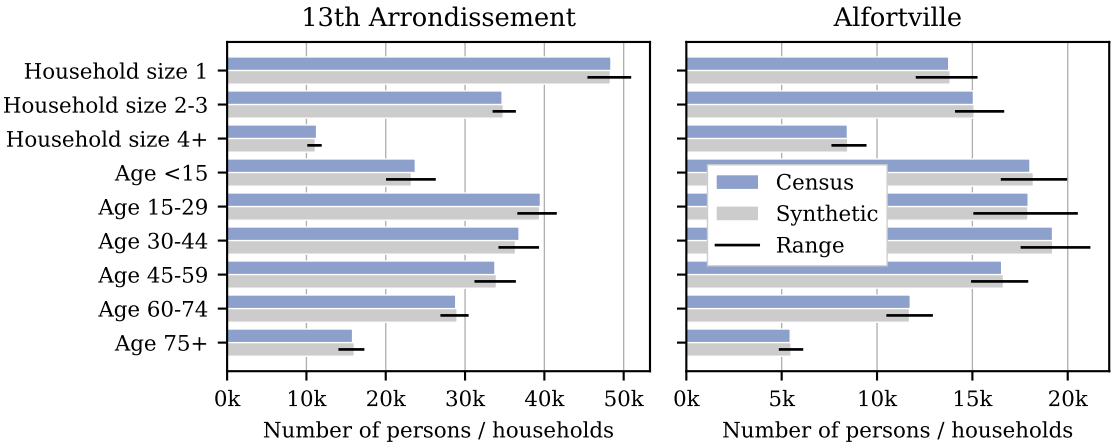


Figure 10: Comparison of sociodemographic attributes between the 13th arrondissement of Paris and the suburban Alfortville municipality.

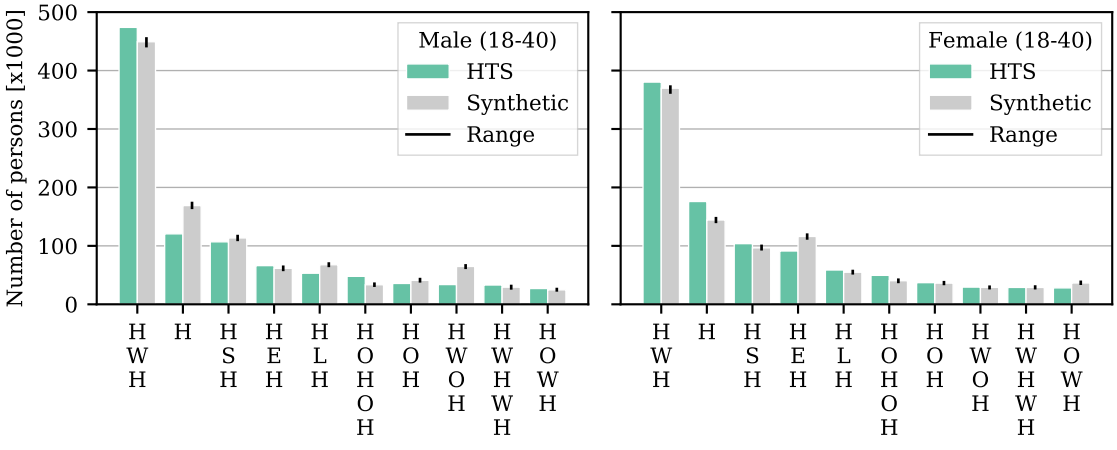


Figure 11: Comparison of the ten most frequent activity chains between the synthetic travel demand and reference data for males and females, respectively, between 18 and 40 years old.

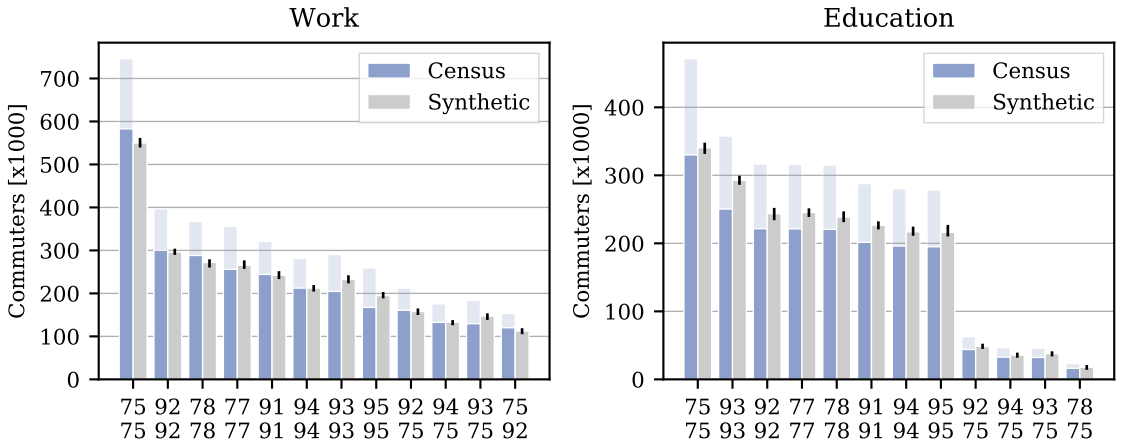
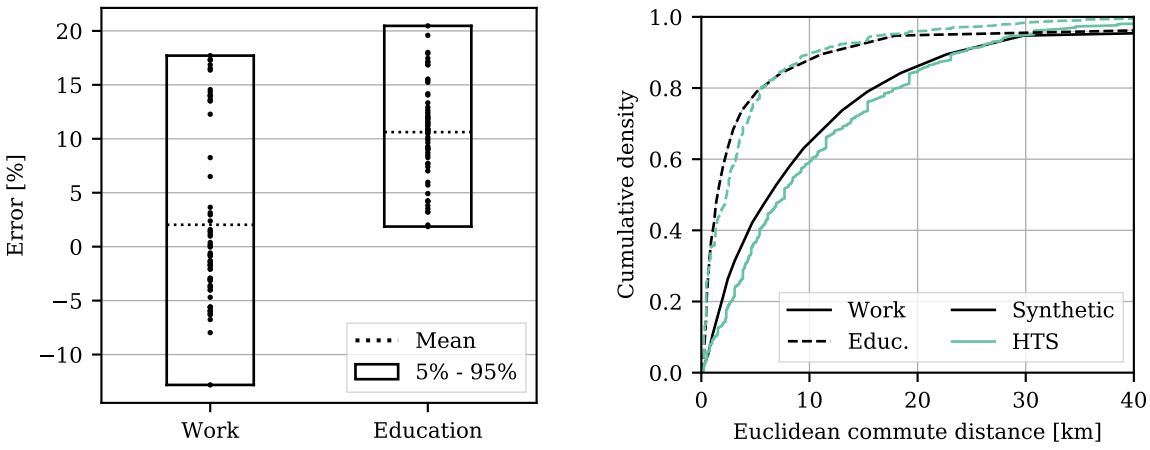


Figure 12: Comparison of top 12 commuting flows between reference data and synthetic population for the departments of Île-de-France. The light bars show the actual reference values from the census while the dark bars indicate a value scaled down by the share of actively commuting persons per day derived from the household travel survey. The top number shows the identifier of the origin department, while the bottom number indicates the destination.

While the previous analyses have shown that the pipeline produces good results in terms of agent attributes, it is also interesting to analyze the synthetic daily plans of the agents. The underlying idea is that activity chains are correlated with sociodemographic attributes and that by applying the matching process, chains are distributed in a meaningful way when analyzed for different population strata. Figure 11 shows the frequencies of the ten most frequent activity chains in the HTS and the synthetic population for people between 18 and 40 years old. The chains for men and women are shown side by side to demonstrate the correlation with population strata. First, Figure 11 clearly shows that starting at home, going to work, and coming back in the evening (HWH) is the most prominent activity chain for both men and women. However, the number of observations for women is less than the frequency for men. On the contrary, staying at home (H) is substantially more frequent for women. The chain distribution for women is more heavy-tailed than for men (in total, less observations fall into the top 10 chains), indicating the phenomenon of stronger trip-chaining, which has been researched in literature (Scheiner and Holz-Rau, 2017).



(a) Modeling error of all departmental commuting flows in Île-de-France.

(b) Comparison of the commuting distance distribution between the synthetic population and reference data.

Figure 13: Comparison of commuting patterns in Île-de-France between the synthetic popualtion and reference data.

After the matching of attributes and activity chains to the synthetic persons, primary locations are assigned to them. These primary locations mainly constitute the commuting patterns of the persons. The validation of those patterns is not straightforward. As the primary reference, we consider the census data set, which gives the home municipality and work/education municipality for each weighted person.

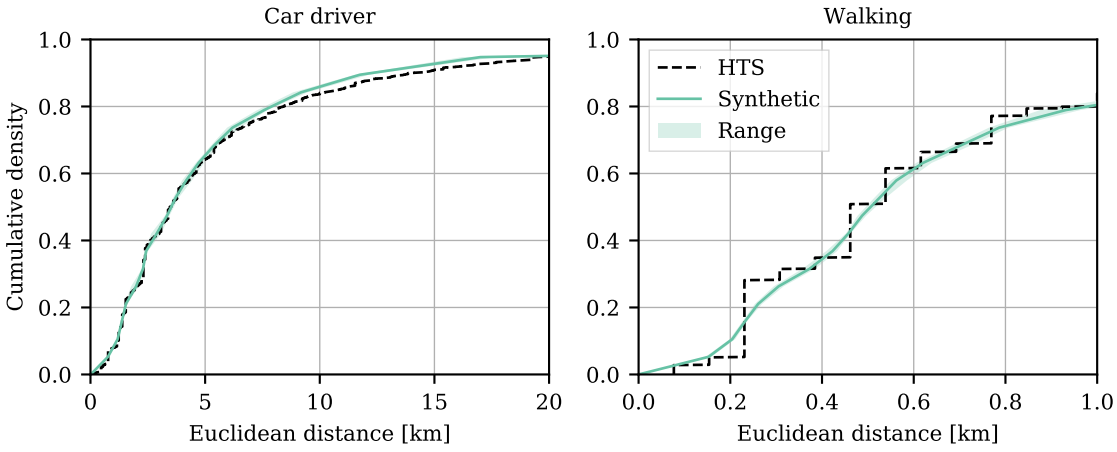


Figure 14: Distance distribution from and to secondary activities by mode. Here, the distributions for car and bicycle are shown as an example. While variation over multiple random seeds is small for *car*, it is noticeable for *bicycle*.

Unfortunately, the synthetic travel demand cannot be easily compared to these values as, on a particular day, not all persons perform their typical commute. Figure 12 shows this discrepancy: The light bar shows commute flows from the census data set, while the gray bar shows the population commute flow between the departments of Île-de-France. Therefore, a correction is applied, leading to the dark bars, for which the synthetic values show a good fit. This correction has been performed by using data from the HTS. From the HTS, we know the departments of the home and work activities of a person. It is then possible to count for each home department k how many (weighted) persons there are that perform a work activity and denote the value as $N_{\mathrm{k,hasActivity}}$. Furthermore, it is known for each person whether she or he is employed, leading a count of $N_{\mathrm{k,isEmployed}}$ for each department k. Finally, a correction factor

$$f_k = N_{\mathrm{k,hasActivity}} / N_{\mathrm{k,isEmployed}} \tag{9}$$

can be obtained, which quantifies the share of employed persons that perform their commute trip on an average weekday. Interestingly, this value is around 0.8 for all departments. A more thorough analysis of the origin of this number would be interesting as it may be affected by the shares of holidays throughout the year, the share of home office during the week, and many other factors. Applied to the commute flows from the census data, new values arise, which can be seen in Figure 12. The procedure for education flows is similar, except that an "is in education" flag from the HTS is used.

Figure 13a shows the distribution of errors between the corrected reference flows and the flows of the synthetic travle demand. The synthetic flows overrepresent the reference value by around 2.5% for work and by around 10% for education on average. However, it remains to debate how stable the process of obtaining the reference values is in itself. Validation with a third data set such as GSM or GPS mobility traces would be a valuable addition in terms of verifying the correction procedure and the synthesized commuting patterns.

Figure 13b shows the distribution of commute *distances* rather than flows between the departments. While the general distribution shape is matched well, errors exist. Further improving the fit of the commuting distance distribution, also separating it by sociodemographic groups, could be a valuable future improvement.

Finally, the synthesis process assigns secondary locations to the activity chains. Figure 14 shows the resulting distance distributions for the *car* and *walking* modes from, to and between secondary activities up to 20km and 1km, respectively. While the first distribution follows the reference distribution from the HTS closely, the *walking* distribution also does so approximately, though large rounding effects in the reference data are apparent.

5. Analysis of sampling rates

In the present part we want to look in more detail at how well the population represents the reference data. As we have seen, it is possible to generate an infinite number of different synthetic travel demand configurations. We call each of them a *realization*. In these realizations we can measure certian quantities, for instance the number of people at a certain age, or the number of commuters between two zones.

Let such a quantity be described by the random variable X. Since there is a distribution of demand configurations, we can examine statistical properties of X such as the expected value $\mathbf{E}[X]$. It is then possible to compare how the sampling rate s and the resulting travel demand configurations X_s lead to different expectation values, confidence bounds or variances $\mathbf{Var}[X_s]$.

Given a reference value $y \in \mathbb{R}$ for the quantity X_s we need to take into account the sampling rate s to make the values comparable. For that, we can define the upscaled quantity $X'_s = X_s/s$. The relative error is then defined as

$$\Delta_s = \left| \frac{X'_s - y}{y} \right| = \left| \frac{X_s}{ys} - 1 \right| \tag{10}$$

for which equally statistics can be calculated such as the expected offset from zero $\mathbf{E}[\Delta_s]$. Using this notation we can quantify the probability of the error being smaller than a predefined threshold $\psi \in \mathbb{R}$ as $\mathbf{P}[\Delta_s \leq \psi]$. By defining a confidence level $\alpha \in [0, 1]$ we can then find the smallest sampling rate s^* at which we are still sufficiently confident that the error stays bounded under a certain threshold:

$$s^* = \sup\{ \quad s \quad | \quad \mathbf{P}[\Delta_s \leq \psi] \geq \alpha \quad \} \tag{11}$$

For instance, we may want that the error in a certain stratum is less than one percent ($\psi = 0.01$) and we want to be sufficiently sure of that. We, therefore, require that in 90% of the popualations we generate ($\alpha = 0.9$) this requirement should be met. Starting at a low sampling rate, for instance at 0.1% we can then successively increase the sampling rate until we find the one where the requirement is met.

This procedure assumes that higher sampling rates generally lead to better fit with reference data. While this makes intuitive sense and has been repeatedly shown for the analysis of population synthesis algorithms, it must be said that this is a hypothesis which could, perhaps, be analysed in more detail for specific synthesis algorithms.

In the analyses in the previous section we generate 20 travel demand configurations and show the sample mean and range for the relevant quantities in comparison to reference data. Here, we want to quantify the influence of the sampling rate on the quality of the output. The sampling rate strongly determines the runtime of the synthesis process, but even more so the runtime of the downstream agent-based transport simulation. Analyzing at which level of downscaling one can maintain a certain level of confidence in the simulation results while minimizing runtime with smaller sample sizes is, therefore, an important task.

In practice, we can not work with the infinitely large distribution of travel demand configurations that can be generated by the algorithms. Rather, we need to stick with a finite sample of size N. It is also not possible to work with the inherent statistical properties of the distribution, but estimates are needed. Using the sample of size N, we can calculate the Monte Carlo estimate of the expectation as

$$\mathbf{E}[X] \approx \hat{\mathbf{E}}[X] = \frac{1}{N} \sum_{i=1}^{N} x_i \ . \tag{12}$$

Likewise, the empirical distribution function of the demand realizations in the sample of size N can be used to estimate quantiles of the underlying distribution of X. To be able to draw conclusions from these estimates it is important to verify that they become sufficiently stable with N samples. Such an analysis is shown in Figure 15a for the number of employed persons between 45 and 59 years old. The visualization shows the reference value (dashed) and the error of $\pm 1\%$ (dotted). The green areas represent the range between the 5% and 95% quantiles estimated from a demand sample of size N. The lightest is estimated from a sample generated with sampling rate 0.1%, the middle one for 1%, and the darkest one for 5%. All intervals stabilize rather quickly after only a few samples. A sample size of 200 is therefore sufficient to perform our analysis. One can see that the light confidence interval (for 0.1% sampling rate) is the widest with a range of almost $\pm 100,000$ persons. It clearly exceeds the error bounds. The darkest interval (for 5% sampling rate) stays within this requirement.

Using the finite demand sample, it is furthermore possible to estimate the probability of the error being smaller than $\psi = 1\%$ given a sampling rate s:

$$\hat{\mathbf{P}}[\Delta_s] = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I} \left[\left| \frac{x_i}{y} - 1 \right| \leq \psi \right] \tag{13}$$

This probability is shown in Figure 15b. Again, the number of employed persons is shown, but furthermore distinguished by certain age strata. For the rather large stratum of people between 45-59

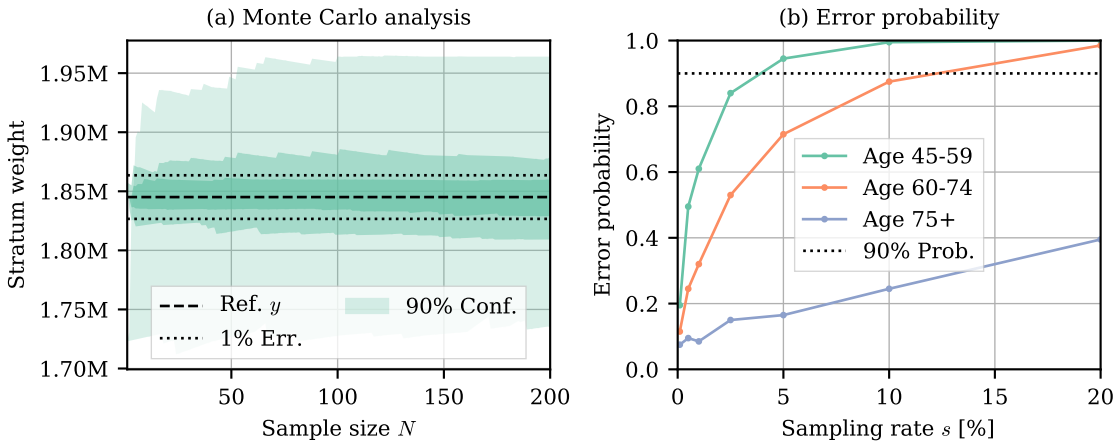


Figure 15: (a) Monte carlo estimates of persons in the stratum "45-59 years old and employed" with 90% confidence interval of the mean with sampling rates of 0.1%, 1%, and 5% (shaded area); and (b) Probability of observing a stratum weight of employed people in a certain age range with an error of less than 1% compared to the reference, estimated on 200 random population realizations.

		Sampling rate s						
		0.001	0.005	0.01	0.025	0.05	0.1	0.2
Variable	Stratum							
Age	<15	0.18	0.41	0.63	0.83	**0.94**	**0.99**	**1.00**
	15-29	0.29	0.46	0.62	0.85	**0.94**	**1.00**	**1.00**
	30-44	0.28	0.50	0.66	0.88	**0.97**	**1.00**	**1.00**
	45-59	0.22	0.53	0.69	0.84	**0.98**	**1.00**	**1.00**
	60-74	0.23	0.45	0.62	0.85	**0.96**	**0.99**	**1.00**
	75+	0.15	0.41	0.59	0.81	**0.93**	**0.99**	**1.00**
Employed	Unemployed	0.39	0.67	0.85	**0.96**	**1.00**	**1.00**	**1.00**
	Employed	0.41	0.72	0.81	**0.97**	**0.99**	**1.00**	**1.00**
Sex	Female	0.42	0.71	0.89	**0.98**	**0.99**	**1.00**	**1.00**
	Male	0.36	0.65	0.85	**0.99**	**1.00**	**1.00**	**1.00**
Socioprof. Cat.	Agriculture	0.00	0.06	0.01	0.06	0.06	0.12	0.21
	Independent	0.07	0.15	0.24	0.43	0.60	0.82	**0.95**
	Science	0.21	0.46	0.58	0.86	**0.94**	**0.99**	**1.00**
	Intermediate	0.17	0.45	0.52	0.82	**0.94**	**0.99**	**1.00**
	Employee	0.15	0.43	0.59	0.85	**0.94**	**0.99**	**1.00**
	Worker	0.23	0.28	0.40	0.67	0.85	**0.95**	**1.00**
	Retired	0.25	0.47	0.68	0.89	**0.99**	**1.00**	**1.00**
	Other	0.28	0.47	0.71	**0.90**	**0.98**	**1.00**	**1.00**
Studies	No	0.46	0.78	**0.95**	**1.00**	**1.00**	**1.00**	**1.00**
	Yes	0.20	0.47	0.64	0.88	**0.97**	**1.00**	**1.00**

Table 3: Probability of observing a deviation of less than 1% from the reference for various population strata, estimated on 200 random population realizations.

years, a confidence of 90% is reached already at a sampling rate between 2.5% and 5%. For the generated demand to represent the stratum from 60-74 years already 10% are necessary. The stratum of employed persons over 75 years, finally, is so small that a smapling rate of 20% can not sufficiently represent this category. In only 40% of generated demand configurations their stratum weight falls well into an error range of ±1% compared to reference data.

The example shows that it is important to think thoroughly about the choice of the sampling rate, even if using small values may seem tempting to save computation time. While a policy study that affects mainly people around 50 years old may be representative for a sampling rate of 10%, a study that focuses on the elderly above 75 is clearly not.

An exhaustive analysis of the generated travel demand in this article would be out of scope and always

incomplete, as the decision which dimensions to analyse for representativeness is highly related to the respective use case. Rather, we want to provide a couple of further examples of this analysis in Table 3. For various population attributes and strata, it shows the probability of the stratum weight to be no further than 1% from their respective reference value. Combinations for which the requirement is fulfilled in more than 90% of the cases are emphasized in bold.

Table 3 shows that in many cases low sampling rates are sufficient, but as soon as smaller quantities should be estimated with confidence (such as small population strata) larger samples are necessary. As this procedure can be repeated with a quantity that can be measured from the synthetic populations (or the downstream agent-based simulation), it is a valuable tool to quantify the representativeness of the synthesis and simulation results.

Finally, it should be mentioned that here we look at the necessary sampling rates to find *one* travel demand configuration that is likely within predefined error bounds. It is often the case that transport simulations are run with varying random seeds and then averaged over a couple of different realizations. The pipeline makes it possible to do so, but with including the whole travel demand synthesis process into this sampling process. It is, therefore, also a valid question to ask what happens if multiple travel demand outcomes are averaged. Generally, the confidence on the outcome (Figure 15a) would be stronger and a specific sampling rate (Figure 15b) is likely to yield a higher confidence when an average is used. In future work it would be interesting to address the trade-off between, for instance, running one simulation with a 100% sampling rate, or 10 simulations with a 5% sampling rate. As simulation times often scale worse than linearly with the demand size, it is a highly relevant question.

6. Discussion

After presenting the synthetic travel demand and the synthesis process, a couple of aspects around the proposed approach should be discussed. First, Section 6.1 provides a critical assessment of various methodological aspects and remarks on using the population in practice. Section 6.2 outlines the future pathway for using the synthetic travle demand in an agent-based transport simulation. In Section 6.3 we cover future ways of extending and improving the synthesis process by including new methods and data sets.

6.1. Methodology

From a methodological perspective, the approach can be discussed quantitatively and qualitatively. From our analysis it becomes evident that the simple data-driven algorithms used do not provide guarantees in generating *consistent* (in the sense of "converging to an unbiased value") outputs. Algorithms need to be specifically designed for that, and while the population sampling procedure seems to have this property (without proof), it is not the case for the statistical matching procedure, when the frequency of activity chains or uncontrolled attributes are examined. From a scientific perspective, it is certainly interesting to follow the topic of designing those algorithms accordingly and checking other approaches (such as Bayesian networks) in that regard. From an applied transport planning perspective, the tools at hand can allow researchers and planners to assess if the generated population is "close enough" for their specific use case.

It is furthermore interesting to think qualitatively about the output of the pipeline. By documenting every step of the synthesis process in detail, we also document the flow of information and how various dimensions of the synthetic persons and their mobility patterns are linked to each other. It is a question that usually cannot be answered when population synthesis is only briefly covered as part of a larger case study: Given the algorithms and data sets used, which questions are valid to answer with the output data? While for every use case, a careful assessment of the correlation structure is important, we want to provide a couple of examples which may not always be obvious, but might become important for the planner:

- Activity chains in the generated population correlate strongly with the sociodemographics of the persons. The sociodemographics of the persons are representative of where these people live. However, it is wrong to assume that there is a direct correlation between types of activity chains and certain regions. The correlation is implicit and *not* controlled.
- On the contrary, income is strongly related to *where* people live. However, in the current state, it is *not* correlated to the sociodemographics of people or households. While again there may be an implicit correlation because certain people live in certain areas, an analysis of how many people, in a certain age group, in an area live in households with a certain income are to be performed with caution and ideally in comparison to a third-party data set.

- Another example is the correlation between income and commuting patterns. It certainly may be possible that income influences the length of commuting trips or the probability of certain destination areas given a fixed origin. However, the way that commute destinations are synthesized in the current process does *not* explicitly model a correlation between destinations and income of the traveler. Hence, an analysis of such a phenomenon also needs to be performed with care.

In general, we would propose to accompany any policy study with reference data that shows that the baseline correlations produced by the synthetic travel demand are realistic. Only then it is possible to add changes to the system and observe effects reliably.

6.2. Transport simulation

While the paper at hand presents the method to create a synthetic travel demand for Île-de-France, the ultimate goal is to perform simulations of the transport system. An example thereof has already been presented by Hörl et al. (2019), where a fleet of automated taxis is simulated in Paris. The following paragraphs will briefly outline the missing steps of setting up a full transport simulation. This topic will be covered in detail in a future publication to allow for extended analysis and discussion of results.

On the way to a full transport simulation, a step of extending the synthetic *travel demand* to a synthetic *transport scenario* needs to be added to the pipeline, namely synthesizing the supply side of the transport system. For Île-de-France, the relevant open data sets can easily be publicly accessed.

OpenStreetMap (*OSM*) is a collaborative project that provides free and open access to geospatial data around the world. It is managed by the OSM Foundation, which has seen rapid growth in the past decade. By now, the project counts around five million registered users creating a world-wide open database of roads, public transport infrastructure, and other points of interest (POIs) represented by more than five billion nodes (OSM, 2020).

To create an open transport scenario, OSM is an invaluable source, because it allows us to use a fine-grained road network that is available in high quality in most parts of the world, especially in Western Europe. Since the whole OSM data set is rather large, it is convenient to make use of smaller regional cut-outs. Fortunately, Geofabrik (2020) provides such a smaller data set for the Île-de-France region.

While the road (and railway) infrastructure is known from OSM, the service does not provide any information on the public transit schedule. Such information can be obtained from a GTFS feed, which is published online by Île-de-France Mobilités (IDFM). The data set is updated weekly and contains information about the location of all public transit stops in the region, the available lines, and their respective schedules.

Currently, the pipeline code includes the capabilities of creating the supply data needed to run simulations in the agent-based transport simulation framework MATSim (Horni et al., 2016). Both OSM and GTFS data sets can be fed into the open source *pt2matsim* tool (Poletti, 2016), which transforms them into a road and rail network, as well as a public transit schedule, that can be used by MATSim.

Several data sets can be used to validate the quality of the simulation output. Some of those are openly available in Île-de-France:

- Traffic count data is a standard source of information used to validate transport simulation results. It allows comparing route choices of individuals on a link level. The regional transport administration offers annual hourly averages, and monthly daily average vehicle counts for the highways in Île-de-France outside of Paris. The city of Paris provides a frequently updated data set of disaggregate hourly vehicle counts on all major arterial roads inside of the city.

- RATP provides tap-in data for their bus and metro services in Île-de-France. This data set can be used to validate transit mode choices (i.e., bus vs. subway) and to validate spatial and temporal mode-choice differences. This data set only contains information for those holding one of the subscription cards and does not capture those travelers without subscription tickets, which limits the usefulness to some extent.

Making use of these data, we can establish an integrated pipeline from raw data sets to a full runnable MATSim simulation. While in this paper, we focus on the synthesis of the demand side, we want to emphasize that the code is available in the current version of the pipeline code to create a full runnable MATSim simulation, which generates a number of further advantages. MATSim simulations are stochastic, i.e., they are only deterministic given a configurable random seed. Hence, it is advised to run multiple MATSim simulations to assess a specific policy's impact. While previously such a sensitivity analysis could only be performed based on a fixed realization of the underlying synthetic travel demand, it can now be performed from raw data to the final policy output. Additionally, it will be possible to

reliably replicate results from downstream MATSim simulations that use a population generated from our approach.

Given the vast amount of studies available for MATSim, our synthetic travel demand opens the possibility to perform similar research on Île-de-France and Paris. Those include analyses on the performance of automated vehicles (Hörl et al., 2019; Bischoff and Maciejewski, 2016; Maciejewski and Bischoff, 2017), their interplay with public transport (Sieber et al., 2020); large-scale adoption of Mobility-as-a-service schemes (Becker et al., 2020a), mitigation of noise and pollutant emissions (Kaddoura and Nagel, 2019; Kaddoura et al., 2020); reliability, interruption management (Leng and Corman, 2020), and optimization of the transit network (Becker et al., 2020b). The ability of MATSim to simulate public transport in detail could be especially interesting for studying the impact of the Grand Paris Express[7], one of the largest infrastructure projects in Europe.

Potential use cases beyond the field of transportation exist. Some examples are evacuation in case of disasters (Lämmel et al., 2009), disease spreading (Hackl and Dubernet, 2019; Muller et al., 2020), or facility location optimization (Balać and Ciari, 2014).

The general pipeline has been adapted for other use cases around the world, such as Switzerland (Hörl et al., 2019), Sao Paulo (Sallard et al., 2020), San Francisco, and Los Angeles (Balac and Hörl, 2021), which shows the flexibility of the proposed approach. Furthermore, all presented data sets, except the regional household travel survey, are available everywhere in France. This means that synthetic travel demand can be created for any city in France, which has already been exemplified with models for the Occitanie region and Toulouse, Nantes, Lyon, and Lille.

While the authors are affiliated and familiar with MATSim as a simulation tool, we intend to pose the pipeline as a simulator-agnostic tool that can serve as an input to various simulators. First experiments with importing the synthetic travel demand into SUMO (Lopez et al., 2018) have already been performed with success.

6.3. Future improvements

There are several paths forward for the agent-based population for Île-de-France. While the version presented here is representing most of the transport demand in the region, there is still room for improvement. As will be explained below, most of these depend on the availability of the data.

- Long-distance commuters to Île-de-France, who are either commuting with high-speed rail or private cars, can be added to enhance the synthetic travel demand. While the share of these travelers is small compared to the local population, it is still valuable to model them as some policy measures, or case studies might heavily depend on those travelers. Here ENTD, traffic counts and information from SNCF can help to improve the model.

- Paris is one of the most popular touristic destinations in the world. On some days, tourists increase the population up to 10%, which affects mostly the public transport network, but also taxi, ride-hailing, and bike-sharing services. The inclusion of tourists in agent-based models was, to the knowledge of the authors, never attempted on a detailed scale. Therefore, it should be one of the major future work efforts, as tourism has become an important aspect in many cities.

- While a manifold of analyses are possible with the synthetic travel demand presented here, results of downstream simulations will become more robust if commercial agents are included as well. For France, several surveys have been performed, such as the Urban Delivery Survey (*Enquête marchandises en ville*) which exists for Île-de-France and other Metropolitan areas and covers logistics movements between companies. Furthermore, there is the Household Delivery Survey (*Enquête Achats Découplés des Ménages*) covering specifically deliveries to households. Unfortunately, they are not published as open data. They are mainly used to feed the statistical model FRETURB (Toilier et al., 2018), which, among other outputs, provides origin-destination matrices for commercial traffic across France. Those could be disaggregated to be used in agent-based transport models. Other methods for the synthesis of freight and commercial travel demand exist, but are often adapted to very specific data sets. With the open enterprise census one common requirement is fulfilled in the French context. On the contrary, knowledge about commrecial transport trips, or even shipments, is not easily accessible (as is common in most places around the world). Relatively generic models for shipment generation, which only rely on enterprise information, exist (Nuzzolo

[7] https://www.societedugrandparis.fr/

and Comi, 2014) and have shown good transferability to other use cases (Nuzzolo et al., 2016). Furthermore, generic models for commercial tour formation have been proposed (e.g. Thoen et al., 2020). Those models could be applied to French data sets to include commercial traffic in a future version of the demand synthesis process.

- The secondary location assignment procedure currently does not take into account the attractiveness of the locations. This fact can lead to over/under-estimation of the number of individuals performing activities at particular shopping or leisure locations. The attractiveness of the place could be measured by the size of the shop/leisure place or its capacity. Such data could be obtained from either OSM or other sources, which would ensure that shopping malls or large supermarkets are attracting mode individuals than smaller shops. A first improvement would be to make use of employment count of distinct shopping or leisure facilities derived from the enterprise census.

- Data sources used in the creation of the synthetic travel demand come from different years, which may have impact on the final results. Census and origin-destination commute flows are from the year 2015, whereas the HTS is from the year 2010. Being the least up to date data set, the HTS does not include some of the changes in the mobility of the population that are visible today. These are mainly an increase in cycling, reduction of car use, or the start of operation of ride-hailing services. It might be relevant to use data sets that are collected within a few years from each other in order to maintain consistency. However, this is not always possible, as, for instance, the ENTD survey is currently performed every ten years. With the speed of change of transportation options, either new sources of people's behavior are needed or an increase in the frequency of general household travel surveys.

- Parking is arguably one of the strongest policy tools, but one rarely used. The difficulties of finding a parking spot in the city of Paris, and its high cost, are among the reasons that have contributed to achieving a very low mode-share of car trips in the city of Paris, which is around 10%. Therefore, modeling parking supply and demand, cost and availability, can be crucial to realistically model individuals' mode-choices. Unfortunately, the data on parking is very scarce, and to our knowledge, there are no publicly available data sources capturing parking. Further exchanges with government agencies and academics in the region, will be important to overcome this limitation and to improve the model.

- Currently, in the process of matching activity chains to individuals, household structure or interaction is not considered. For the EGT, each household member was interviewed. Therefore, it would be possible in the matching phase to maintain the interactions that exist within the household. This would ensure that joint trips in households are modeled properly, and that shared resources like cars or bicycles can be distributed appropriately for use throughout the day, and among household members. This would ensure, for instance, that if a household owns one car, it can only be driven by one member of the household at a time. The documentation of the lack of household and social interactions in models and some proposed solutions can be found in Arentze and Timmermans (2009); Ronald et al. (2012).

The modularity and flexibility of the travel demand synthesis pipeline allow for straightforward additions of new modules, models, and data sets. Therefore, all the above can be included in an improved Île-de-France population, as long as the data is available. In terms of validation, it will be an important next step to directly compare the outcomes of the pipeline when used with the regional household survey EGT in comparison to the national ENTD which is available to the public.

7. Conclusion

We have presented a generalized travel demand synthesis framework for agent-based transport simulation. The framework is open-source and allows building populations that are entirely reproducible by others. The framework was used here to generate a synthetic travel demand of the Île-de-France region in France. While at this stage, we use simple algorithms to create the population, it should serve as a benchmark for further improvements. Nevertheless, the population was validated and can serve as the starting point for transportation studies and beyond.

While we present the process for the Île-de-France region, the pipeline has already been used to create synthetic travel demand for Switzerland, Sao Paulo, San Francisco, and Los Angeles, which are all, except Switzerland, based on open data. This shows the flexibility and modularity of the approach.

Furthermore, the process can be repeated for any city or region in France, which has already been done for a number of places such as Lille, Lyon, Toulouse, and Nantes.

Going beyond bare synthesis, we also conduct an error analysis on the generated population depending on the sampling rate. To our knowledge, this is the first effort to quantify the error that scaling down the population can have on the outcome and possible results of the studies.

We also highlighted the necessity of being aware of correlations between certain characteristics of the population that exist in practice and the generated population. This has previously been widely neglected in agent-based studies. However, these correlations between different population characteristics can be crucial to accurately asses the impacts of transportation policies studied with agent-based models.

Finally, the framework can be easily extended with new, more complex algorithms, and thus the population can be improved. Access to other data sources can also further improve the process. By making the pipeline open-source, it is the hope of the authors that other researchers and practitioners will be inclined to contribute, both their coding skills and data.

Acknowledgement

We would like to acknowledge **Airbus Urban Mobility GmbH** whose funding has supported the development of an agent-based simulation of the Île-de-France region around Paris.

Appendix A. Glossary

BD-TOPO French address database

BPE Service and facility database for France (*Base permanente des équipements*)

DRIEA Regional representation of the ministry of environment and energy (*Direction régionale et interdépartementale de l'Équipement et de l'Aménagement*)

EGT Regional household travel survey fir Île-de-France (*Enquête globale de transport*)

ENTD National household travel survey for France (*Enquête national transports et déplacements*)

FiLoSoFi Income tax database for France (*Fichier Localisé Social et Fiscal*)

GTFS General Transit Feed Specification

HTS Household travel survey

IDFM Regional public transport authority in Île-de-France (*Île-de-France Mobilités*)

IGN National geographic institute of France (*Institut national de l'information géographique et forestière*)

INSEE National Institute of Statistics and Economic Studies, the national statistical office in France (*Institut national de la statistique et des études économiques*)

IRIS Smallest statistical zoning system in France (*Îlots Regroupés pour l'Information Statistique*)

OD Origin-destination, for instance for commuting data

OSM OpenStreetMap

RATP Public transport provider in Paris (*Régie Autonome des Transports Parisiens*)

RP Popuation census for France (*Recensement de la population*)

SIRENE Enterprise census of France (*Système national d'identification et du répertoire des entreprises et de leurs établissements*)

SNSF Swiss National Science Foundation)

STIF Former public transport authority of Île-de-France (now IDFM) (*Syndicat des Transports d'Île-de-France*)

Appendix B. Technical documentation

All code that is referenced in this paper (version `1.1.0`) is available online at:

`https://github.com/eqasim-org/ile-de-france/tree/v1.1.0`

The repository provides instructions that make it possible for anybody with basic knowledge of Python to create a synthetic travel demand as discussed above. Three steps need to be followed:

- **Collection of data sets**: We do not provide the output data of the pipeline, but anybody can regenerate it. For that, the individual raw source data sets need to be collected. The repository gives instructions on where to find the data sets, how to download them, and how to arrange them as input data for the pipeline. All of them are publicly accessible.
- **Running travel demand synthesis**: As the next step, instructions are provided on how to run the demand synthesis. For that, *Python* needs to be installed. We explain how to install all needed dependencies in a *conda*[8] environment, how to apply minimal necessary adjustments to the pipeline configuration (setting the path to the input data), and running the code. The output is the travel demand data, as described below.
- **Run a MATSim simulation**: *Optionally*, the pipeline provides the tools to create the input data for a full MATSim (Horni et al., 2016) simulation. We provide instructions on how to generate this data and run a full simulation of the Île-de-France region. This step requires *Java* and *osmosis*[9].

The output data of the travel demand synthesis is structured as follows:

- `meta.json` contains some metadata, e.g., with which random seed or sampling rate the population was created and when.
- `persons.csv` and `households.csv` contain all persons and households in the population with their respective sociodemographic attributes.
- `activities.csv` and `trips.csv` contain all activities and trips in the daily mobility patterns of these people including attributes on the purposes of activities or transport modes for the trips.
- `activities.gpkg` and `trips.gpkg` represent the same trips and activities as spatial data that can be processed in geographical information systems. Activities contain point geometries to indicate where they happen, and the trips file contains line geometries to indicate the origin and destination of each trip.

For the tables, *CSV* was chosen as the format because it is universally known and easy to use. *GeoPackage* (GPKG) was chosen for the spatial data because, contrary to ESRI shapefiles, the data is provided as a single file, and the format reproduces the same descriptive, full-length column names as in the CSV files. At the same time, it provides projection information, contrary to other formats such as *GeoJSON*. All spatial data is provided in the `EPSG:2154` projection that is commonly used for France.

[8] `https://www.anaconda.com/`
[9] `https://github.com/openstreetmap/osmosis`

References

ActivitySim, 2020. An open platform for activity-based travel modeling. URL: https://activitysim.github.io/. Accessed 27 Apr 2020.

Adnan, M., Pereira, F.C., Azevedo, C.M.L., Basak, K., Lovric, M., Raveau, S., Zhu, Y., Ferreira, J., Zegras, C., Ben-Akiva, M., 2016. Simmobility: A multi-scale integrated agent-based simulation platform, in: 95th Annual Meeting of the Transportation Research Board Forthcoming in Transportation Research Record.

Ahrens, A., Lyons, S., 2020. Do rising rents lead to longer commutes? A gravity model of commuting flows in Ireland. Urban Studies , 1–16.

Arentze, T., Timmermans, H., Hofman, F., 2007. Creating Synthetic Household Populations: Problems and Approach. Transportation Research Record: Journal of the Transportation Research Board 2014, 85–91.

Arentze, T.A., Timmermans, H.J., 2009. A need-based model of multi-day, multi-person activity generation. Transportation Research Part B: Methodological 43, 251 – 265.

Auld, J., Hope, M., Ley, H., Sokolov, V., Xu, B., Zhang, K., 2016. Polaris: Agent-based modeling framework development and implementation for integrated travel demand and network and operations simulations. Transportation Research Part C: Emerging Technologies 64, 101 – 116.

Auld, J., Mohammadian, A., 2009. Framework for the development of the agent-based dynamic activity planning and travel scheduling (ADAPTS) model. Transportation Letters 1, 245–255.

Axhausen, K.W., Gärling, T., 1992. Activity-based approaches to travel analysis: conceptual frameworks, models, and research problems. Transport reviews 12, 323–341.

Baker, M., 2016. 1,500 scientists lift the lid on reproducibility. Nature 533.

Balać, M., Ciari, F., 2014. Retailers location choice based on shopping and land prices, in: 21st International Conference on Recent Advances in Retailing and Services Science, EIRASS.

Balac, M., Hörl, S., 2021. Synthetic population for the state of California based on open-data: examples of San Francisco Bay area and San Diego County, in: 100th Annual Meeting of the Transportation Research Board, Washington, D.C., January 2021.

Ballis, H., Dimitriou, L., 2020. Revealing personal activities schedules from synthesizing multi-period origin-destination matrices. Transportation Research Part B: Methodological 139, 224–258.

Becker, H., Balac, M., Ciari, F., Axhausen, K.W., 2020a. Assessing the welfare impacts of shared mobility and mobility as a service (maas). Transportation Research Part A: Policy and Practice 131, 228 – 243.

Becker, H., Manser, P., Hörl, S., Axhausen, K.W., 2020b. Designing a large-scale public transport network using agent-based microsimulation. Transportation Research Part A: Policy and Practice .

Bhat, C., Guo, J., Srinivasan, S., Sivakumar, A., 2008. CEMDAP User's Manual. Center for Transportation Research, University of Texas 3.

Bischoff, J., Maciejewski, M., 2016. Simulation of city-wide replacement of private cars with autonomous taxis in berlin. Procedia computer science 83, 237–244.

Bonabeau, E., 2002. Agent-based modeling: Methods and techniques for simulating human systems. Proceedings of the National Academy of Sciences 99, 7280–7287.

Borysov, S.S., Rich, J., Pereira, F.C., 2019. How to generate micro-agents? A deep generative modeling approach to population synthesis. Transportation Research Part C: Emerging Technologies 106, 73–97.

Bösch, P.M., Müller, K., Ciari, F., 2016. The ivt 2015 baseline scenario. 16th Swiss Transport Research Conference .

Boulton, G., 2016. International accord on open data. Nature 530, 281–281.

Cacioppo, J.T., Kaplan, R.M., Krosnick, J.A., Olds, J.L., Dean, H., 2015. Social, behavioral, and economic sciences perspectives on robust and reliable science .

Chapuis, K., Taillandier, P., Renaud, M., Drogoul, A., 2018. Gen*: a generic toolkit to generate spatially explicit synthetic populations. International Journal of Geographical Information Science 32, 1194–1210.

Chen, X., Dallmeier-Tiessen, S., Dasler, R., Feger, S., Fokianos, P., Gonzalez, J.B., Hirvonsalo, H., Kousidis, D., Lavasa, A., Mele, S., et al., 2019. Open is not enough. Nature Physics 15, 113–119.

Diogu, W.O., 2019. Towards the Implementation of an Activity-based Travel Demand Model for Emerging Cities: Integrating TASHA and MATSim. Master's thesis.

D'Orazio, M., Di Zio, M., Scanu, M., 2006. Statistical matching: Theory and practice. John Wiley & Sons.

Erath, A., Fourie, P.J., van Eggermond, M.A., Ordonez Medina, S.A., Chakirov, A., Axhausen, K.W., 2012. Large-scale agent-based transport demand model for singapore. Arbeitsberichte Verkehrs-und Raumplanung 790.

Esztergár-Kiss, D., Rózsa, Z., Tettamanti, T., 2018. Extensions of the Activity Chain Optimization Method. Journal of Urban Technology 25, 125–142.

Esztergár-Kiss, D., Rózsa, Z., Tettamanti, T., 2020. An activity chain optimization method with comparison of test cases for different transportation modes. Transportmetrica A: Transport Science 16, 293–315.

Farooq, B., Bierlaire, M., Hurtubia, R., Flötteröd, G., 2013. Simulation based population synthesis. Transportation Research Part B: Methodological 58, 243–263.

Fournier, N., Christofa, E., Akkinepally, A.P., Azevedo, C.L., 2020. Integrated population synthesis and workplace assignment using an efficient optimization-based person-household matching method. Transportation .

Garrido, S., Borysov, S.S., Pereira, F.C., Rich, J., 2020. Prediction of rare feature combinations in population synthesis: Application of deep generative modelling. Transportation Research Part C: Emerging Technologies 120, 102787.

Geofabrik, 2020. Ile-de-france. URL: `https://download.geofabrik.de/europe/france/ile-de-france.html`. Accessed 24 Apr 2020.

Goodman, S.N., Fanelli, D., Ioannidis, J.P.A., 2016. What does research reproducibility mean? Science Translational Medicine 8. doi:`10.1126/scitranslmed.aaf5027`.

Gupta, S., Agrawal, A., Gopalakrishnan, K., Narayanan, P., 2015. Deep learning with limited numerical precision. arXiv 1502.02551.

Hackl, J., Dubernet, T., 2019. Epidemic spreading in urban areas using agent-based transportation models. Future Internet 11, 92.

Hao, J.Y., 2009. TASHA-MATSim integration and its application in emission modelling. Master's thesis.

He, B.Y., Zhou, J., Ma, Z., Chow, J.Y., Ozbay, K., 2020. Evaluation of city-scale built environment policies in New York City with an emerging-mobility-accessible synthetic population. Transportation Research Part A: Policy and Practice 141, 444–467.

Hörl, S., Axhausen, K.W., 2020. Relaxation-discretization algorithm for spatially constrained secondary location assignment, in: 99th Annual Meeting of the Transportation Research Board, Washington, D.C., January 2020.

Hörl, S., Becker, F., Dubernet, T.D., Axhausen, K.W., 2019. Induzierter Verkehr durch autonome Fahrzeuge: Eine Abschätzung, Schlussbericht, SVI 2016/001. Schriftenreihe 1650, UVEK, Bern.

Horni, A., Nagel, K., Axhausen, K.W., 2016. The Multi-Agent Transport Simulation MATSim. Ubiquity Press, London.

Hörl, S., Balac, M., Axhausen, K.W., 2019. Dynamic demand estimation for an AMoD system in Paris, in: 30th IEEE Intelligent Vehicles Symposium, Paris, June 2019.

INSEE, 2003. Professions et catégories socioprofessionnelles. URL: https://www.insee.fr/fr/metadonnees/pcs2003/categorieSocioprofessionnelleAgregee/1. Accessed 24 Apr 2020.

INSEE, 2010. Île-de-france à la page: Les franciliens consacrent 1 h 20 par jour à leurs déplacements.

Jensen-Butler, C., 1972. Gravity Models as Planning Tools: A Review of Theoretical and Operational Problems. Geografiska Annaler: Series B, Human Geography 54, 68–78.

Joubert, J.W., de Waal, A., 2020. Activity-based travel demand generation using Bayesian networks. Transportation Research Part C: Emerging Technologies 120, 102804.

Justen, A., Martínez, F.J., Cortés, C.E., 2013. The use of space–time constraints for the selection of discretionary activity locations. Journal of Transport Geography 33, 146–152.

Kaddoura, I., Bischoff, J., Nagel, K., 2020. Towards welfare optimal operation of innovative mobility concepts: External cost pricing in a world of shared autonomous vehicles. Transportation Research Part A: Policy and Practice 136, 48–63.

Kaddoura, I., Nagel, K., 2019. Congestion pricing in a real-world oriented agent-based simulation context. Research in Transportation Economics 74, 40 – 51.

Kamel, J., Vosooghi, R., Puchinger, J., Ksontini, F., Sirin, G., 2018. Exploring the Impact of User Preferences on Shared Autonomous Vehicle Modal Split: A Multi-Agent Simulation Approach. Transportation Research Procedia 37, 115–122.

Kickhofer, B., Hosse, D., Turnera, K., Tirachinic, A., 2016. Creating an open matsim scenario from open data: The case of santiago de chile. http://www. vsp. tuberline. de/publication: TU Berlin, Transport System Planning and Transport Telematics .

Kitamura, R., 1988. An evaluation of activity-based travel analysis. Transportation 15, 9–34.

Lämmel, G., Klüpfel, H., Nagel, K., 2009. The matsim network flow model for traffic simulation adapted to large-scale emergency egress and an application to the evacuation of the indonesian city of padang in case of a tsunami warning. Pedestrian behavior , 245–265.

Leng, N., Corman, F., 2020. The role of information availability to passengers in public transport disruptions: An agent-based simulation approach. Transportation Research Part A: Policy and Practice 133, 214–236.

Levenshtein, V.I., 1966. Binary codes capable of correcting deletions, insertions, and reversals, in: Soviet physics doklady, pp. 707–710.

Liu, F., Janssens, D., Cui, J., Wets, G., Cools, M., 2015. Characterizing activity sequences using profile Hidden Markov Models. Expert Systems with Applications 42, 5705–5722.

Lopez, P.A., Behrisch, M., Bieker-Walz, L., Erdmann, J., Flötteröd, Y.P., Hilbrich, R., Lücken, L., Rummel, J., Wagner, P., Wießner, E., 2018. Microscopic Traffic Simulation using SUMO, in: 21st IEEE International Conference on Intelligent Transportation Systems, IEEE.

Lovelace, R., 2020. Open access transport models: A leverage point in sustainable transport planning. Transport Policy , 8.

Lovelace, R., Ballas, D., 2013. 'truncate, replicate, sample': A method for creating integer weights for spatial microsimulation. Computers, Environment and Urban Systems 41, 1 – 11.

Ma, T.Y., Klein, S., 2018. Bayesian networks for constrained location choice modeling using structural restrictions and model averaging , 21.

Maciejewski, M., Bischoff, J., 2017. Congestion effects of autonomous taxi fleets , 1–10.

Martinez, L.M., Correia, G.H.A., Viegas, J.M., 2015. An agent-based simulation model to assess the impacts of introducing a shared-taxi system: an application to Lisbon (Portugal): AN APPLICATION TO LISBON (PORTUGAL). Journal of Advanced Transportation 49, 475–495.

Martinez, L.M., Viegas, J.M., 2017. Assessing the impacts of deploying a shared self-driving urban mobility system: An agent-based model applied to the city of Lisbon, Portugal. International Journal of Transportation Science and Technology 6, 13–27.

Muller, S.A., Balmer, M., Neumann, A., Nagel, K., 2020. Mobility traces and spreading of covid-19. medRxiv .

Müller, K., 2017. A Generalized Approach to Population Synthesis. Ph.D. thesis. ETH Zurich.

Nuzzolo, A., Comi, A., 2014. Urban freight demand forecasting: A mixed quantity/delivery/vehicle-based model. Transportation Research Part E: Logistics and Transportation Review 65, 84–98.

Nuzzolo, A., Comi, A., Ibeas, A., Moura, J.L., 2016. Urban freight transport and city logistics policies: Indications from Rome, Barcelona, and Santander. International Journal of Sustainable Transportation 10, 552–566.

OSM, 2020. Stats. URL: `https://wiki.openstreetmap.org/wiki/Stats`. Accessed 24 Apr 2020.

Pendyala, R.M., Bhat, C.R., Goulias, K.G., Paleti, R., Konduri, K.C., Sidharthan, R., Hu, H.H., Huang, G., Christian, K.P., 2012. Application of Socioeconomic Model System for Activity-Based Modeling: Experience from Southern California. Transportation Research Record: Journal of the Transportation Research Board 2303, 71–80.

Poletti, F., 2016. Public transit mapping on multi-modal networks in MATSim. Master Thesis. IVT, ETH Zurich, Zurich.

Rasouli, S., Timmermans, H., 2014. Activity-based models of travel demand: promises, progress and prospects. International Journal of Urban Sciences 18, 31–60.

Recker, W.W., 1995. The household activity pattern problem: general formulation and solution. Transportation Research Part B: Methodological 29, 61–77.

Rich, J., Mulalic, I., 2012. Generating synthetic baseline populations from register data. Transportation Research Part A: Policy and Practice 46, 467–479.

Ronald, N., Arentze, T., Timmermans, H., 2012. Modeling social interactions between individuals for joint activity scheduling. Transportation Research Part B: Methodological 46, 276 – 290.

Saadi, I., Eftekhar, H., Teller, J., Cools, M., 2018a. Investigating scalability in population synthesis: a comparative approach. Transportation Planning and Technology 41, 724–735.

Saadi, I., Farooq, B., Mustafa, A., Teller, J., Cools, M., 2018b. An efficient hierarchical model for multi-source information fusion. Expert Systems with Applications 110, 352–362.

Saadi, I., Mustafa, A., Teller, J., Cools, M., 2016a. Forecasting travel behavior using Markov Chains-based approaches. Transportation Research Part C: Emerging Technologies 69, 402–417.

Saadi, I., Mustafa, A., Teller, J., Farooq, B., Cools, M., 2016b. Hidden Markov Model-based population synthesis. Transportation Research Part B: Methodological 90, 1–21.

Sallard, A., Hörl, S., Balac, M., 2020. Agent-based scenario of Sao Paulo Metropolitan Area. Working paper. IVT, ETH Zurich, Zurich.

Scheiner, J., Holz-Rau, C., 2017. Women's complex daily lives: a gendered look at trip chaining and activity pattern entropy in germany. Transportation 44, 117–138.

Shoval, N., Isaacson, M., 2007. Sequence Alignment as a Method for Human Activity Analysis in Space and Time. Annals of the Association of American Geographers 97, 282–297.

Sieber, L., Ruch, C., Hörl, S., Axhausen, K.W., Frazzoli, E., 2020. Improved public transportation in rural areas with self-driving cars: A study on the operation of swiss train lines. Transportation Research Part A: Policy and Practice 134, 35–51.

Stark, P.B., 2018. Before reproducibility must come preproducibility. Nature 557, 613–614.

STIF, OMNIL, DRIEA, 2013. Enquête globale transport: La ville de paris.

Sun, L., Erath, A., 2015. A Bayesian network approach for population synthesis. Transportation Research Part C: Emerging Technologies 61, 49–62.

Sun, L., Erath, A., Cai, M., 2018. A hierarchical mixture modeling framework for population synthesis. Transportation Research Part B: Methodological 114, 199–212.

Thoen, S., Tavasszy, L., de Bok, M., Correia, G., van Duin, R., 2020. Descriptive modeling of freight tour formation: A shipment-based approach. Transportation Research Part E: Logistics and Transportation Review 140, 101989.

Toilier, F., Gardrat, M., Routhier, J., Bonnafous, A., 2018. Freight transport modelling in urban areas: The French case of the FRETURB model. Case Studies on Transport Policy 6, 753–764.

Viegas, J.M., Martínez, L.M., 2010. Generating the universe of urban trips from a mobility survey sample with minimum recourse to behavioural assumptions, in: Proceedings of the 12th World Conference on Transport Research.

Vitins, B.J., Erath, A., Axhausen, K.W., 2016. Integration of a Capacity-Constrained Workplace Choice Model: Recent Developments and Applications with an Agent-Based Simulation in Singapore. Transportation Research Record: Journal of the Transportation Research Board 2564, 1–13.

Vosooghi, R., Kamel, J., Puchinger, J., Leblond, V., Jankovic, M., 2019a. Robo-taxi service fleet sizing: Assessing the impact of user trust and willingness to use. Transportation 46, 1997–2015.

Vosooghi, R., Puchinger, J., Jankovic, M., Vouillon, A., 2019b. Shared autonomous vehicle simulation and service design. Transportation Research Part C: Emerging Technologies 107, 15 – 33.

Yoon, S.Y., Deutsch, K., Chen, Y., Goulias, K.G., 2012. Feasibility of using time–space prism to represent available opportunities and choice sets for destination choice models in the context of dynamic urban environments. Transportation 39, 807–823.

Ziemke, D., Kaddoura, I., Agarwal, A., 2019a. Entwicklung eines regionalen, agentenbasierten verkehrssimulationsmodells zur analyse von mobilitätsszenarien für die region ruhr, in: Mobilität in Zeiten der Veränderung. Springer, pp. 383–410.

Ziemke, D., Kaddoura, I., Nagel, K., 2019b. The matsim open berlin scenario: A multimodal agent-based transport simulation scenario based on synthetic demand modeling and open data. Procedia computer science 151, 870–877.

Ziemke, D., Nagel, K., Bhat, C., 2015. Integrating CEMDAP and MATSim to increase the transferability of transport demand models. Transportation Research Record 2493, 117–125.

Île-de-France Mobilités, OMNIL, DRIEA, 2010. Enquête globale transport 2010.