Introducing the eqasim pipeline

From raw data to agent-based transport simulation

Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 184 (2021) 712–719

Procedia Computer Science

www.elsevier.com/locate/procedia

The 10th International Workshop on Agent-based Mobility, Traffic and Transportation Models, (ABMTRANS) March 23 - 26, 2021, Warsaw, Poland

Introducing the *eqasim* pipeline: From raw data to agent-based transport simulation

Sebastian Hörl[a], Milos Balac[b,*]

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

Abstract

This paper introduces the *eqasim* framework with the aim to provide a consistent pipeline from raw data to a final transport simulation. It therefore lays the foundation to achieve fully reproducible agent-based transport simulations. While the pathway from raw data to a generic synthetic travel demand was covered previously for specific use cases, here the general methodology is summarized. Furthermore, the tools and methods for combining MATSim simulations with flexibly definable discrete choice models is described, which is the core of the existing simulation implementations of *eqasim* for Île-de-France, Switzerland, Sao Paulo and California.

Keywords: agent-based; transport; simulation; reproducibility; open data; open source; eqasim

1. Introduction

Agent-based models have found their way in the field of transportation. Their ability to model interactions of travelers with one another and the environment, together with the rising computational power of modern computers has led to emergence of several agent-based transport simulation frameworks such as MATSim [11], SUMO [15] and others.

Initial work in the field of agent-based modeling in transportation was mainly focused on showcasing the importance and benefits of these models, while constantly improving their capabilities. However, little effort was put in documenting and ensuring reproducibility of the conducted studies.

Each case study conducted with an agent-based model mainly relies on two fundamental blocks: (1) synthetic travel demand and transport supply as input data, and (2) the simulation environment. The simulation environment

* Corresponding author. Tel.: +41-44-633-37-30.
E-mail address: milos.balac@ivt.baug.ethz.ch

1877-0509
10.1016/j.procs.2021.03.089

can be made accessible to others by publishing the code open-source, providing good documentation and maintaining version control. This ensures the minimum requirements for any of the studies to be reproducible. However, access to the synthetic travel demand and the tools to reproduce it are also essential parts leading to the repeatability of agent-based modeling studies.

To foster reproducibility of agent-based studies, we propose a general pipeline called *eqasim* that provides a clear path from raw data to a final agent-based mobility simulation.

The paper is structured as follows. Section 2 gives a brief overview of agent-based simulation and models. Section 3 provides more information on the *eqasim* approach. Section 4 showcases one implementation of the pipeline with a model for Zurich, Switzerland. Section 5 provides discussion and concluding remarks.

2. Background

Agent-based models in transportation arrived as the need to model individual travelers and their interaction with the environment was rising. Their importance was evident when traditional transport models struggled to evaluate the complex interactions of users and shared mobility services i.e., car-sharing. While car-sharing was probably one of the first emerging transport services to be investigated in detail using agent-based models [16, 4], shared on-demand automated services ensured that agent-based models become a tool of choice to deal with this kind of transportation services (i.e. [13, 6]).

The significance of agent-based models led several research groups to develop agent-based modeling frameworks (MATSim [11], POLARIS [2], SimMobility [3], TRANSIMS [17], SUMO [15]). There are hundreds of studies conducted using these models that gave insights for many relevant topics - equity, policy, disease spreading, welfare analysis, and many others. However, the complexity of these models makes them hard and time consuming to set-up, which is one of their important limitations [14]. Once the model is set up the studies can be conducted. Yet, these studies are almost always decoupled from the process that leads to the generation of the necessary input, which makes them difficult to repeat by other researchers. This, therefore, prevents reproducibility of the majority of agent-based modeling studies, which is the second important limitation identified in the current literature on agent-based models [14].

Hence, there is a need to provide the tools that allow users to replicate and repeat research on agent-based studies. We propose a framework called *eqasim* that takes raw data, creates a synthetic travel demand data set, allows the user to convert the data into an agent-based model (currently MATSim) and to simulate traveler behavior.

3. From raw data to an agent-based simulation

That scientific research should be easily accessible and repeatable are the guiding principles of the *eqasim* methodology. We aim to provide a clear path from raw data to final agent-based simulation that is easily extendable, modifiable, and verifiable. Ideally, all elements on this path should be published as open-source code and data should be open and publicly available as well. This way, it would be possible for anybody interested to gather the publicly available data, run the code and reproduce a synthetic travel demand and mobility scenario that has been used in research elsewhere.

Such a process has many advantages. First, research becomes reproducible and results can be verified. While this should be the standard, it is often not possible when it comes to agent-based transport simulation. The same applies to applied planning projects, which could be performed in a more transparent way if the entire process of setting up the required simulations were open.

The path from the raw data to the mobility simulation (here referred to as the *pipeline*) is divided in three parts in (see Figure 1):

- Population pipeline
- Converter
- Agent-based simulation

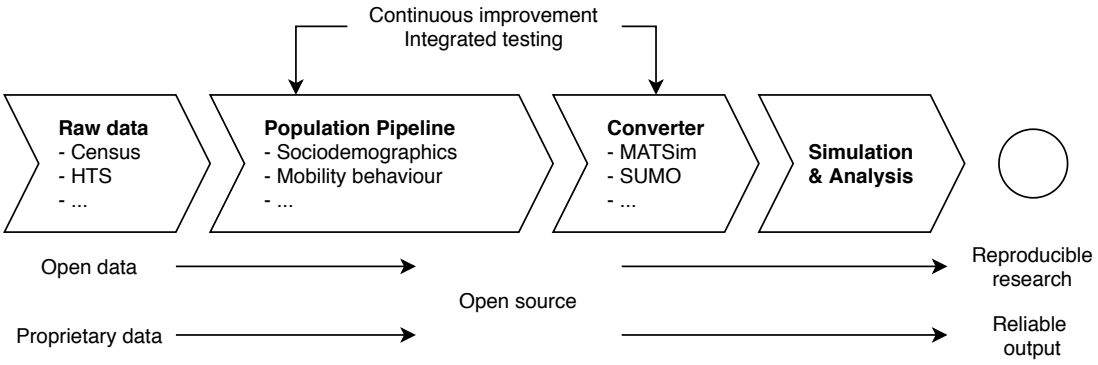

Fig. 1. Eqasim pipeline

The backbone of the pipeline is `synpp`[1], a generic Python package for chaining algorithms and code pieces (*stages*) in a larger pipeline set-up. While it can be used in a more general context, it aims at providing a solid basis for travel demand synthesis and transport simulation applications.

The *Population Pipeline* part of pipeline takes raw data, performs all the necessary filtering and cleaning steps, and produces the synthetic population with activity patterns and activity locations. This process is ideally based on the openly available datasets published by government agencies: household travel survey, population census, enterprise census, commuting flows, etc. An example of generating a synthetic travel demand completely based on open data was previously described in [12], where the Île-de-France region in France is taken as the case study. The paper describes the complete process, including all methods used, and points to the open-source code that can be used to reproduce the synthetic travel demand.

The output of this step contains the following data-sets:

- *households.csv* that contains all households in the study region characterized with certain attributes, e.g., household size, household type, car and/or bike ownership, household income
- *persons.csv* that contains all individuals living in the study area characterized with further attributes, e.g., age, gender, public season ticket ownership
- *trips.csv* and *trips.gpkg* that contain information on all trips conducted by the individuals in plain and geolocalized form
- *activities.csv* and *activities.gpkg* that contain information on all activities performed by individuals during an average day in plain and geolocalized form.

The second step in the process is to convert the synthetic travel demand to the right format for the use in an agent-based model. Currently the main implementation is a converter for MATSim [11], but first experiments with converting the data into input for SUMO have been conducted.

The third and the final step is running the mobility simulation, which will be presented in more detail in the following. *eqasim* provides the environment to easily integrate the converted travel demand from the second part of the pipeline into the MATSim simulation, though we add further extensions to increase compatibility with other existing concepts methods in transport planning.

MATSim simulations normally consist of three phases: mobility simulation, replanning and scoring. The recently added Discrete Mode Choice (DMC)[2] [7, 8] module, however, allows to completely replace the scoring procedure with discrete mode-choice models that are executed during the replanning stage. To date, all models developed with the *eqasim* framework make use of this component to simulate mode choice decisions.

[1] https://github.com/eqasim-org/synpp
[2] https://github.com/matsim-org/matsim-libs/tree/master/contribs/discrete_mode_choice

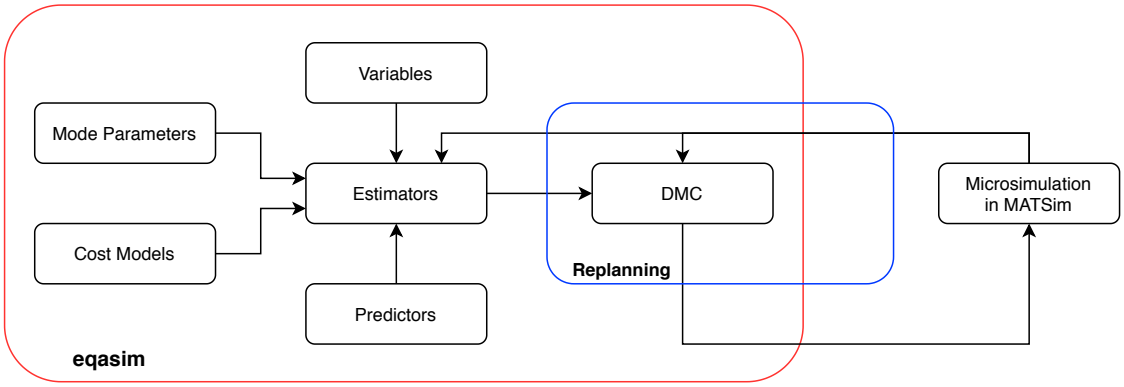

Fig. 2. High-level structure of the *eqasim* simulation

To ease setting up a simulation with a discrete mode choice model, the *eqasim-java* package has been developed, which is available as open-source[3]. Besides easy-to-use interfaces and tools to set up highly customizable discrete choice models in MATSim it provides further utilities for routing trips, cutting scenarios, and more, which are compatible with and used in the overall *eqasim* pipeline.

The main components of the *eqasim-java* package are (Figure 2):

- **Variables:** Each alternative and person can have specific attributes stored that can be later used by utility estimators (e.g., mobility tool ownership, household income, ...).
- **Cost models:** Each mode alternative has an implementation of the associated cost model. The cost can depend on the traveled distance or time, attributes of the individuals, area where the trip is happening, etc.
- **Mode-choice parameters:** For each alternative, mode-choice parameters can be defined via an input file or the command line. For non-default parameters, extension of the `ModeParameters` class is possible. The standard model implementation contains a wide range of variables describing the most important decision factors for the modes *car*, *public transport*, *bicycle*, and *walking*.
- **Predictors:** Each alternative has its own predictor that calculates expected travel time, waiting time, transfer time, and access/egress time, if relevant for a specific mode. The information is obtained for each trip and based on information provided by the routers implemented in MATSim. The information for each trip and for each alternative is cached to reduce computation time.
- **Estimators:** each mode alternative has an estimator, which calculates its utility, based on the parameters, cost model, predicted trip characteristics, and person attributes. Default estimators are available for the modes mentioned above.

The estimators quantify the utility of each mode choice alternative that is defined in the discrete mode choice extension. The extension then performs the mode choice, usually making use of the calculated utilities as part of a mutlinomial logit model. However, other formulations, such as a nested logit model, are available. The discrete choice model itself is called in the replanning phase of MATSim and applied to a randomly selected share of agents in each iteration, based on a configurable share.

4. An example of Zurich scenario

In the following an example of using the pipeline for a transport simulation of Zurich, Switzerland, is documented. The synthetic travel demand for the study was first used in [10] with an updated version published in [9] where the relevant data sets are documented. The complete list of data-sets and the code to generate the synthetic travel demand

[3] https://github.com/eqasim-org/eqasim-java

are also available online[4]. The pipeline makes use of census data, a detailed household structure survey, a household travel survey and additional data sets. Unfortunately, contrary to the pipeline implementation for Paris and Île-de-France [12], not all data sets are publicly available as open data. For that reason it is currently only possible for Swiss research institutions to completely reproduce the synthetic travel demand.

The transport supply part of the model is generated using the pt2matsim [18] converter. The OpenStreetMap is used to generate the road network, and GTFS schedules are used to generate the public transport services. Finally, pt2matsim is used to map the generated transit schedules to the road netowrk.

4.1. Mode choice model

The mode choice model is a multinomial logit model which was estimated using the Swiss household travel survey and a specific travel survey for the Zurich region. The estimated choice parameters are defined inside the *Mode Parameters* component of the *eqasim* structure (Figure 2). The utility functions of the model are defined in the *Estimators* component. Most inputs to the choice model are static based on a trip characteristics or travelers attributes, but they do not change over multiple iterations. The major influence is the travel time by car as it is directly dependent on the traffic assignment resulting from the detailed vehicle simulation in MATSim. The following equations define the utility functions for the modes *car*, *public transport*, *bicycle*, and *walking*:

$$
\begin{aligned}
u_{car} = & \beta_{\text{ASC,car}} \\
& + \beta_{\text{inVehicleTime,car}} \cdot \xi_{\text{TD}} \cdot x_{\text{inVehicleTime,car}} \\
& + \beta_{\text{work,car}} \cdot x_{\text{work}} + \beta_{\text{city,car}} \cdot x_{\text{city}} \\
& + \beta_{\text{cost}} \cdot \xi_{\text{CD}} \cdot \xi_{\text{CI}} \cdot x_{\text{cost,car}}
\end{aligned}
$$

$$
\begin{aligned}
u_{bicycle} = & \beta_{\text{ASC,bicycle}} \\
& + \beta_{\text{travelTime,bicycle}} \cdot \xi_{TD} \cdot x_{\text{travelTime,bicycle}} \\
& + \beta_{\text{highAge,bicycle}} \cdot [a_{\text{age}} \geq 60]
\end{aligned}
$$

$$
\begin{aligned}
u_{walk} = & \beta_{ASC,walk} \\
& + \beta_{travelTime,walk} \cdot \xi_{TD} \cdot x_{\text{travelTime,walk}}
\end{aligned}
$$

$$
\begin{aligned}
u_{pt} = & \beta_{\text{ASC,pt}} \\
& + \beta_{\text{inVehicleTime,train}} \cdot \xi_{\text{TD}} \cdot x_{\text{inVehicleTime,train}} \\
& + \beta_{\text{inVehicleTime,other}} \cdot \xi_{\text{TD}} \cdot x_{\text{inVehicleTime,other}} \\
& + \beta_{\text{inVehicleTime,feeder}} \cdot x_{\text{inVehicleTime,feeder}} \\
& + \beta_{\text{transferTime,pt}} \cdot x_{\text{transferTime,pt}} \\
& + \beta_{\text{accessEgressTime,pt}} \cdot x_{\text{accessEgressTime,pt}} \\
& + \beta_{\text{headway,pt}} \cdot x_{\text{headway,pt}} \\
& + \sum_{G} \beta_{\text{ptQuality,G}} \cdot x_{\text{ptQuality,G}} \\
& + \beta_{\text{cost}} \cdot \xi_{\text{CD}} \cdot \xi_{\text{CI}} \cdot x_{\text{cost,pt}}
\end{aligned}
$$

All β represent model parameters that were estimated while x represent per-trip input variables and a represent per-agent variables. The utility function for public transport makes a difference between routes that include a train with $x_{\text{inVehicleTime,train}}$ quantifying the travel time in the main stage and $x_{\text{inVehicleTime,feeder}}$ quantifying the rest. Only if no train is included in the route, $x_{\text{inVehicleTime,other}}$ has a value while the other two are set to zero. The *pt quality* parameters and variables are based on a methodology of the Federal Office for Spatial Development quantifying the accessibility by public transport of any location in Switzerland, based on vicinity to public transport infrastructure and frequencies of the accessible transport lines [1].

The model includes two interaction terms ξ which establish non-linear dependencies of the utility of travel time and cost on distance

$$
\xi_{TD} = \left(\frac{x_{\text{euclideanDistance}}}{\theta_{\text{referenceDistance}}} \right)^{\lambda_{\text{TD}}} \qquad \text{and} \qquad \xi_{CD} = \left(\frac{x_{\text{euclideanDistance}}}{\theta_{\text{referenceDistance}}} \right)^{\lambda_{\text{CD}}} \tag{1}
$$

[4] https://gitlab.ethz.ch/ivt-vpl/populations/ch-zh-synpop

Table 1. Mode choice model (Source: [9])

	Parameter	
Private car	$\beta_{\text{ASC,car}}$	0.224
	$\beta_{\text{inVehicleTime,car}}$	-0.019
	$\beta_{\text{work,car}}$	-1.161
	$\beta_{\text{city,car}}$	-0.459
Bicycle	$\beta_{\text{ASC,bicycle}}$	0.152
	$\beta_{\text{travelTime,bicycle}}$	-0.126
	$\beta_{\text{highAge,bicycle}}$	-2.659
Walking	$\beta_{\text{ASC,walk}}$	0.590
	$\beta_{\text{travelTime,walk}}$	-0.046
Others	β_{cost}	-0.089
	μ_{SP}	0.942
	λ_{CI}	-0.817
	λ_{CD}	-0.221
	λ_{TD}	0.115
	$\theta_{\text{referenceDistance}}$	39.000
	$\theta_{\text{referenceIncome}}$	12.260

	Parameter	
Public transport	$\beta_{\text{ASC,pt}}$	0.000
	$\beta_{\text{inVehicleTime,feeder}}$	-0.045
	$\beta_{\text{inVehicleTime,other}}$	-0.012
	$\beta_{\text{inVehicleTime,train}}$	-0.007
	$\beta_{\text{transferTime,pt}}$	-0.012
	$\beta_{\text{accessEgressTime,pt}}$	-0.014
	$\beta_{\text{headway,pt}}$	-0.030
	$\beta_{\text{ptQuality,B}}$	-1.744
	$\beta_{\text{ptQuality,C}}$	-1.641
	$\beta_{\text{ptQuality,D}}$	-0.965
	$\beta_{\text{ptQuality,None}}$	-1.089

and one interaction term that relates household income to the perception of cost:

$$\xi_{CI} = \left(\frac{a_{\text{householdIncome}}}{\theta_{\text{referenceIncome}}} \right)^{\lambda_{\text{CI}}} \tag{2}$$

The estimated mode parameters are documented in Table 1.

The *Cost Model* for *private car* alternative calcualtes the cost of car travel as 0.26 CHF/km based on the routed distance. The *Cost Model* for public transport defines the cost based on the subscription ownership. Fares for public transport are zero if the agent has an annual subscription ("Generalabo"), which is very common in Switzerland; they are also zero if the agent has a regional subscription and the origin and destination of the trip are within 15 km Euclidean distance of his or her home location (model assumption). Otherwise, the fare is calculated as 0.5 CHF/km based on the routed distance inside a public transport vehicle (thus excluding access and egress walks). If the agent has a "Halbtax" half fare subscription, the fare is reduced by half.

4.2. Calibration

Figure 3 shows the results in terms of mode shares by Euclidean distance for the choice model after it was integrated directly into MATSim (in light gray). As usual with this procedure, the resulting shares do not fit perfectly with the reference data obtained from the household travel survey. For that reason, two adjustments needed to be done. First, the survey used to estimate the choice model was not representative for shorter distances. Hence, we do not see a good fit for the walking transport mode. Because of that, the utility for the *walking* mode was modified with an additional penalty term which is close to zero for very short trips and becomes strongly negative (−100) once a certain threshold travel time is reached:

$$u'_{\text{walk}} = u_{\text{walk}} - \exp\left(\log(101) \cdot \frac{x_{\text{travelTime,walk}}}{\theta_{\text{walkThreshold}}} \right) + 1 \tag{3}$$

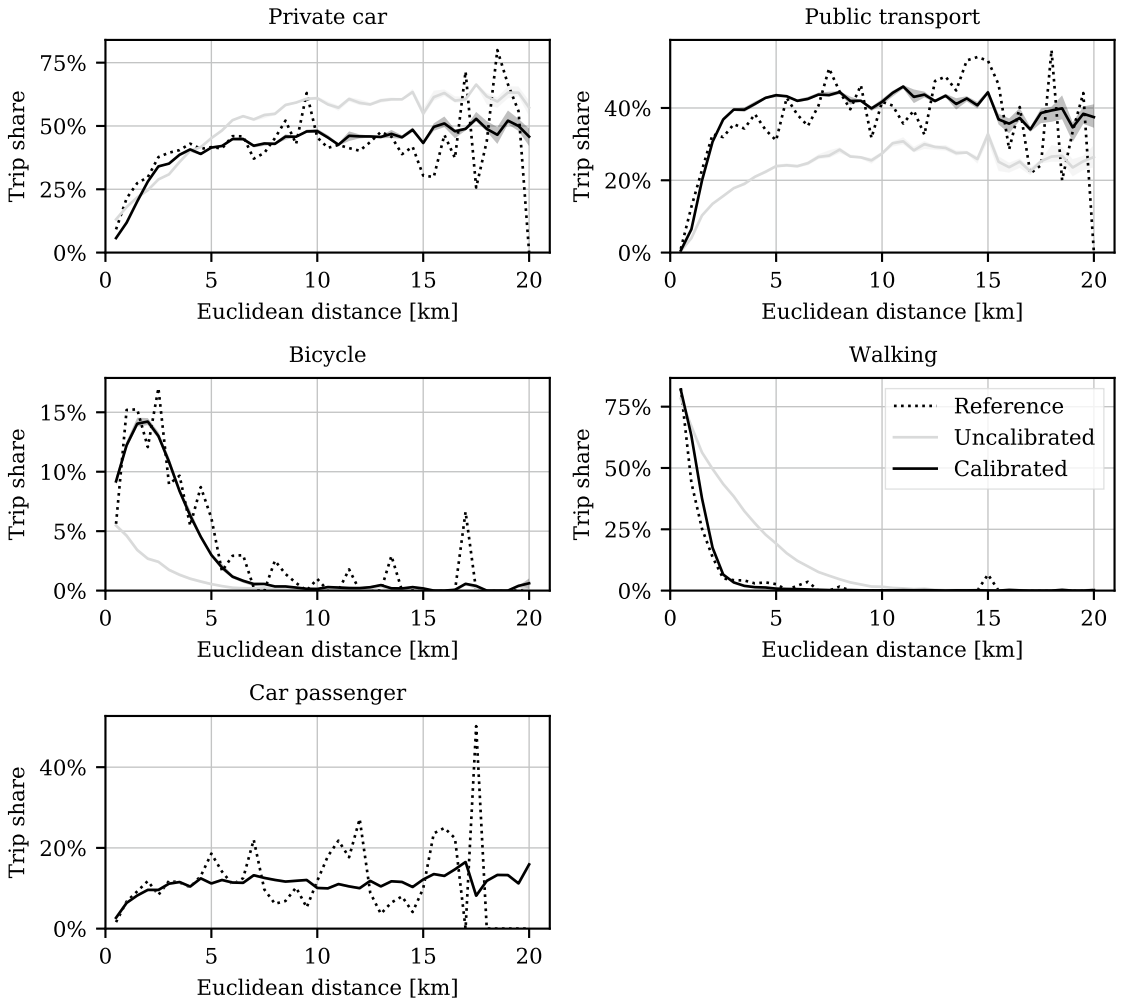

Fig. 3. Mode share calibration of the baseline model (Source: [9])

Second, the alternative-specific constant for the *car* mode was adjusted. This is usual necessary as the interpretation of travel time in the survey data and in the simulation differ. Generally, the alternative-specific constant captures all components of the trip utility which are not described explicitly by the other terms. Hence, it can include the discomfort of parking search, paying for parking, and, in the specific case of this simulation, access and egress to the vehicle. Although recent versions of MATSim support simulating access and egress stages to and from the vehicle, this was not used here, and even then, the model does not include the locations of parking spots or garages, or any additional model of choosing between on-street parking, using a large garage or even parking facilities provided by the company for work trips. Such considerations will be important for the future development of the framework. To compensate for such effects, $\beta_{\text{ASC,car}}$ was set to $\beta'_{\text{ASC,car}} = -0.8$ after several steps of manual calibration.

Finally, Figure 3 shows the fit of the calibrated model to the reference data based on the national travel survey (in black). While the reference data itself is noisy, one can see that the simulated shares follow closely the trend observed in the data.

5. Conclusion

This paper presented the *eqasim* pipeline that takes raw data and leads to an agent-based simulation after a number of sequential processing steps. The Switzerland model is used as an example. The proposed framework, which was already successfully applied to other regions, namely California [5], Sao Paulo [19], and France [12], enables reproducible agent-based transportation studies. Even though the framework itself is not the guarantee of reproducibility of the downstream studies, it provides the users with the necessary tools to achieve it.

The framework by its design is modular, which enables each of the stages to be replaced. Different methods can be used in travel demand synthesis. Different converters can be implemented to convert the demand to the input format for other agent-based models. Eventually, different agent-based models can be employed for the final simulation studies individually or in combination and comparison.

References

[1] ARE, 2011. ÖV-Güteklassen - Berechnungsmethodik ARE. Technical Report. Bundesamt für Raumentwicklung (ARE).
[2] 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.
[3] Azevedo, C.L., Deshmukh, N.M., Marimuthu, B., Oh, S., Marczuk, K., Soh, H., Basak, K., Toledo, T., Peh, L.S., Ben-Akiva, M.E., 2017. Simmobility short-term: An integrated microscopic mobility simulator. Transportation Research Record 2622, 13–23.
[4] Balac, M., Becker, H., Ciari, F., Axhausen, K.W., 2019. Modeling competing free-floating carsharing operators–a case study for zurich, switzerland. Transportation Research Part C: Emerging Technologies 98, 101–117.
[5] 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.
[6] Bischoff, J., Maciejewski, M., Nagel, K., 2017. City-wide shared taxis: A simulation study in Berlin, in: 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC), IEEE. pp. 275–280.
[7] Hörl, S., Balac, M., Axhausen, K.W., 2018. A first look at bridging discrete choice modeling and agent-based microsimulation in MATSim. Procedia Computer Science 130, 900 – 907.
[8] Hörl, S., Balac, M., Axhausen, K.W., 2019a. Pairing discrete mode choice models and agent-based transport simulation with MATSim, in: 98th Annual Meeting of the Transportation Research Board, Washington, D.C.
[9] Hörl, S., Becker, F., Axhausen, K.W., 2021. Simulation of price, customer behaviour and system impact for a cost-covering automated taxi system in Zurich. Transportation Research Part C: Emerging Technologies 123, 102974.
[10] Hörl, S., Becker, F., Dubernet, T.D., Axhausen, K.W., 2019b. Induzierter Verkehr durch autonome Fahrzeuge: Eine Abschätzung, Schlussbericht, SVI 2016/001. Schriftenreihe 1650, UVEK, Bern.
[11] Horni, A., Nagel, K., Axhausen, K.W., 2016. The Multi-Agent Transport Simulation MATSim. Ubiquity Press, London.
[12] Hörl, S., Balac, M., 2020. Open data travel demand synthesis for agent-based transport simulation. A case study of Paris and Île-de-France. Arbeitsberichte Verkehrs- und Raumplanung 1499.
[13] 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.
[14] Kagho, G.O., Balac, M., Axhausen, K.W., 2020. Agent-based models in transport planning: Current state, issues, and expectations. Procedia Computer Science 170, 726–732.
[15] Lopez, P.A., Behrisch, M., Bieker-Walz, L., Erdmann, J., Flötteröd, Y.P., Hilbrich, R., Lücken, L., Rummel, J., Wagner, P., WieBner, E., 2018. Microscopic traffic simulation using sumo, in: 21st International Conference on Intelligent Transportation Systems (ITSC), IEEE. pp. 2575–2582.
[16] Martínez, L.M., Correia, G.H.d.A., Moura, F., Mendes Lopes, M., 2017. Insights into carsharing demand dynamics: Outputs of an agent-based model application to Lisbon, Portugal. International Journal of Sustainable Transportation 11, 148–159.
[17] Nagel, K., Rickert, M., 2001. Parallel implementation of the transims micro-simulation. Parallel Computing 27, 1611–1639.
[18] Poletti, F., 2016. Public transit mapping on multi-modal networks in MATSim. Master Thesis. IVT, ETH Zurich, Zurich.
[19] Sallard, A., Balać, M., Hörl, S., 2020. Synthetic travel demand for the Greater São Paulo Metropolitan Region, based on open data. Under review .