A generalizable pipeline for agent-based transport models in France
Short paper for hEART 2020

Sebastian Hörl, Christoper Tchervenkov, Milos Balac
ETH Zurich, Institute for Transport planning and systems

Abstract

Agent-based simulation is increasingly used in transport modeling and planning as it allows to study in detail emerging new modes of mobility and individual-based policy impacts. An important component of such models are synthetic populations of travellers that represent the travel demand. While various agent-based transport simulation tools have evolved over the past decade, open tools for consistent, reproducible, adaptable and verifiable synthesis of synthetic populations are still scarce. Yet, evolving open data policies all over the world make it now possible to reach this goal. The following describes an open-source pipeline from raw data to final agent-based transport simulation that is based on publicly available data in France. While the discussion focuses on the use case of Île-de-France and Paris, an outline on how the process can be generalized to other French regions is provided. Finally, other use cases such as Sao Paulo, San Francisco and Los Angeles, to which the appraoch has been adopted, are presented.

Introduction

In recent years, research on new mobility services such as on-demand and autonomous mobility has strongly increased. Interest in these services is mainly fostered by increasing digitalization and faster means of communication which make it possible to not only make mobility decisions on the scale of days or hours, but also immediately. Therefore, these services pose new challenges for transport simulation, as detailed interactions between customers, operators and vehicles need to be modeled in order to understand the trade-offs between service levels, service costs and operational strategies.

Therefore, the tool of choice in many research projects is agent-based simulation. Such simulations allow to follow the trajectories and actions of single entities such as travellers and on-demand vehicles on very small time scales with high spatial resolution. Furthermore, a more fine-grained distinction of travellers by, for instance, sociodemographic attributes allow for additional analyses that have not been possible previously.

Various agent-based transport simulators are available, such as MATSim (Horni et al., 2016), SUMO (Lopez et al., 2018) or Polaris (Auld et al., 2016) and applications and use cases are abundant. Yet, many publications only provide limited detail on how the underlying scenario has been generated from data. In the case of MATSim, many scenarios for various cities and regions around the world exist[1], but only a few of them have been thoroughly documented and made public for other researchers to verify. To the knowledge of the

[1] https://matsim.org/gallery/

authors, this is only the case for the MATSim scenarios of Santiago de Chile (Kickhofer et al., 2016) and Berlin (Ziemke et al., 2019).

Hence, our question is how to make research on agent-based transport models reproducible, verifiable and testable. While working on agent-based transport simulations in MATSim over the past years, the authors have developed a flexible open-source scenario synthesis pipeline that has now been applied to over five different cities worldwide with more methodological work on the synthesis process being performed on the example of Île-de-France and Paris. In this case, the main idea is to provide a pipeline that can entirely be based on publicly available, open data and itself is published as open source code. By that, we intend to foster open and reproducible research in agent-based transport models.

The following gives a brief overview of the conducted work as well as hints on how the process can be generalized and used by other researchers. First, we will go into detail about the aims of this research; second, the specific case of Île-de-France will be outlined, followed by a discussion on further simulation scenarios that have recently been developed and closing remarks.

Open and collaborative approach

The central idea of the population/scenario synthesis pipeline that will be outlined further below is to allow researchers to go from raw data to full agent-based transport simulations in one consistent stream of models and transformations (see Figure 1). Ideally, all elements in this process 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 population and mobility scenario that has been used in research elsewhere.

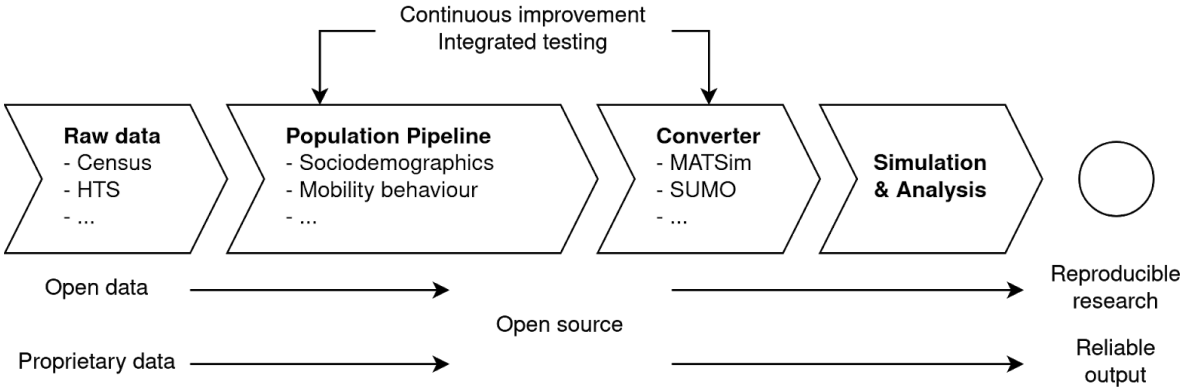

Figure 1. Scenario synthesis process

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.

Second, researchers usually work on rather specific elements of such a pipeline. For instance, there are many algorithms for population synthesis such as the examples described by Sun et al. (2015, 2018) or Saadi et al. (2016). Those publications usually focus on a specific part of the pipeline, yet it is not clear to which extent the algorithms can improve the overall quality of a full-stack transport simulation. Hence, a pipeline as proposed by us would allow researchers to test their algorithms in an integrated environment.

Third, building a mobility scenario can become a rather complex project with many steps and dependencies between models and data sets, which makes it difficult to keep track of changes, intermediate alterations of algorithms, etc. A pipeline as shown in Figure 1 therefore makes it possible to apply integration tests to the pipeline, which will alert the user in case small changes on certain parameters or algorithms lead to drastic changes in the output data.

The pipeline for Île-de-France

As an example, we want to introduce our scenario synthesis pipeline for Île-de-France, which is currently being documented. While here we give a broad overview of the process, Hörl et al. (2020) will describe all steps in detail and perform a thorough analysis of synthesis errors. Furthermore, the paper explores the variance of population attributes across multiple realizations of the synthetic population and how it evolves when downsampling the population to allow for faster simulation. The aim of the pipeline is to start with raw data sets, to transform them, to apply further models, and arrive at a final running agent-based transport simulation. We intentionally use rather simple and straight-forward algorithms in the current state to establish a baseline against which future implementations of more complex and elaborate algorithms can be tested. For simulating we use the agent- and activity-based transport simulation framework MATSim.

The proposed pipeline for Île-de-France consists of a number of steps that each draw from a specific data set and apply a certain algorithm to make use of the data in a synthetic transport scenario.

In France a detailed census data set[2] is made publicly available every year by the statistical office. The data set contains sociodemographic information about persons grouped into households for the whole country. Households and persons are given as single observations, but they can only be geolocalized on the level of statistical zones. Such zones are defined such that each of them contains at least 200 inhabitants to allow for sufficient anonymization. Further anonymization measures are applied to areas with even lower population density. As the data set is not comprehensive, we use the household weights that are provided by the statistical office to scale up the population. It is then possible to generate artificial persons in all Île-de-France with sociodemographic attributes such as the number of vehicles, household size (on the household level), and age, gender, socio-professional

[2] Recensement de la population

category, and employment status (on the person level). Home coordinates are sampled at random within the statistical zone of each household.

Unfortunately, census data does not provide income information, which is important to realistically model mobility behaviour and allows for equity analyses and more, in the final simulations. Therefore, we use the publicly available Filosofi[3] data set, which aggregates tax data for all France and provides decile-based household income distributions for all statistical zones and municipalities. Currently, we use the data to randomly sample household incomes dependent on municipality for each household. In the future, the data set would even allow further refinements as distributions are also provided per household type.

Third, we use household travel surveys to attach activity chains to the synthetic population. For Île-de-France, the pipeline can use the publicly available national household travel survey[4], which, unfortunately, is rather sparse as only less than 5,000 person activity chain observations are available for Île-de-France. Alternatively, the regional household travel survey[5] can be used with around 35,000 activity chains. However, the EGT is not publicly available and must be obtained on request from the involved authorities. Both data sets are rather old with the former being conducted around 2010 and the latter around 2012. It is possible to choose either of the data sets in the pipeline code. The chosen data set is then used in a *statistical matching* procedure to attach a daily activity chain to each of the synthetic agents. This is done by comparing the age, gender, socio-professional category, household income and car ownership attributes from the synthetic agents to the persons in the household travel survey. We then sample one chain dependent on weights given in the HTS from the set of all persons that match those attributes. Activity chains contain all activities that a person did during one day. They are defined by type (currently “home”, “work”, “education”, “shopping”, “leisure” and “other”) and they have start and end times. Furthermore, the daily activity chains contain trips, which connect those activities and are defined by a certain mode of transport.

Fourth, we use origin-destination commuting data which is provided along with the census data and is also publicly available. The flows are given on the level of municipalities and for work and educational commutes. Given the home municipality of each synthetic person, we sample a destination zone from the resulting OD matrix. In the future, this could be further refined by mode of transport as the data allows for that. A specific location is chosen with the help of the BPE[6], which is a publicly available data set on all enterprises in France. Given the destination zone, a random observation is drawn from this data set to assign a specific coordinate to the “work” or “education” activities in a synthetic person’s activity chain. All other activities (*secondary activities*) are assigned the location of an observation from the BPE where, for instance, *shopping,* is available and in such a way that the distance distributions in the synthetic population follow the reference data from the household travel survey. This procedure is described in more detail in Hörl et al. (2020).

[3] Dispositif sur les revenus localisés sociaux et fiscaux
[4] Enquête nationale transports et déplacements
[5] Enquête globale de transport (EGT)
[6] Base permanente des équipements

Finally, the synthetic population is converted to the MATSim format. Additionally, we use an extract of Île-de-France from OpenStreetMap to create a MATSim-format road network and also the digital public transport schedule for Île-de-France is converted to the simulation format. This schedule is made public by the regional public transport authority Île-de-France mobilités. At the end of this pipeline, the simulation can be run. Figure 2 shows a snapshot of such a large-scale agent-based simulation of Île-de-France.

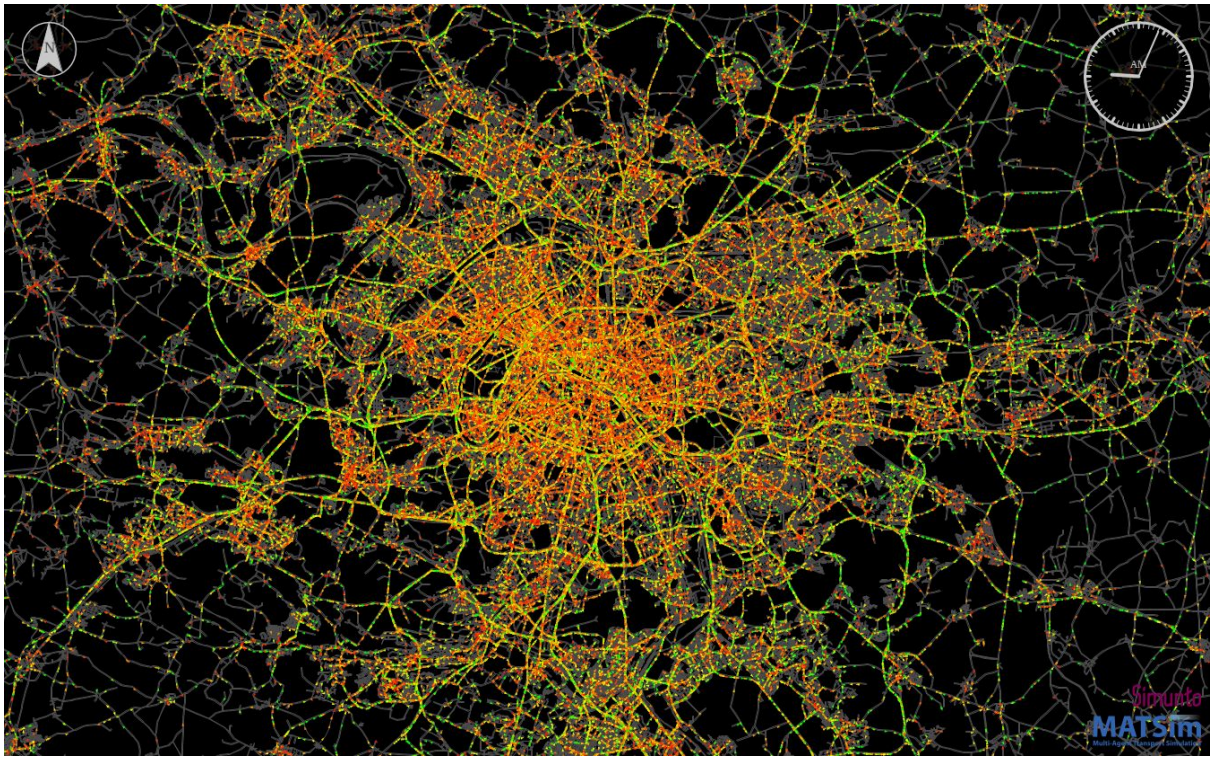

Figure 2: Agent-based MATSim simulation of Île-de-France.

At the time of writing, this population is actively used in ongoing research projects on new mobility services at ETH Zurich and environmental impacts at ENPC in Paris. Furthermore, a case study on simulating a fleet of automated taxis in Paris (Hörl et al., 2019) has recently been conducted based on this population.

Generalizing the pipeline

The approach as outlined above has already been applied for other use cases. The most established scenario is our model of Switzerland, which has been developed for a project on the analysis of automated mobility services (Hörl et al., 2019). Furthermore, a scenario for Sao Paulo (Sallard et al., 2020) is available. While the former cannot be easily shared because census and other data sets are proprietary, the latter can also be built from open and publicly available data. Currently, efforts are being made to apply the same process to Quebec City, Montreal and Jakarta.

Additional scenarios for San Francisco and Los Angeles (Balac, 2020) have been developed using the same pipeline tools. There, we follow the spirit of using the pipeline as an integrative environment as the step of synthesizing households, persons and

sociodemographic attributes is replaced by the population synthesis algorithm PopGen by MARG (2016) that is openly available.

For the case of France, we observe that the pipeline is mainly based on data that is available for the whole country: census, household travel survey (in case the national one is used), tax data, enterprise census, and OpenStreetMap. Only public transport schedules in GTFS or similar format are not available everywhere in the country, and preferably a more detailed and up-to-date household travel survey than the ENTD should be used. Several French cities such as Toulouse and Lyon provide such data sets, meaning that our open source synthesis pipeline could easily be used to create agent-based simulations.

In terms of generalization it should also be mentioned that our current simulations are mainly focused on the MATSim framework. However, only the very last part of converting the synthetic population, the network and transit schedule is actually MATSim-specific. With little effort, it is possible to add different converters and run simulations on other platforms. First tests in that direction have been done with the SUMO simulator.

Conclusion

Our research around agent-based transport models over the past years has led to the development of an open population/scenario synthesis pipeline. The intent is to collect a number of high quality open data transport scenarios that can be used, verified and improved by researchers all over the world. By providing the software in a flexible, extensible way, we invite everybody interested to replace certain components with their own algorithms and techniques and to test them in an integrated transport simulation environment going from raw data to the final simulation.

For the future, the current state of the pipeline allows a number of pathways for research: We see an interesting project in implementing advanced population synthesis algorithms into the pipeline (such as the examples by Sun et al. (2015, 2018) or Saadi et al. (2017)) and comparing their performance, not only in terms of generating “one” population, but also in terms of variance. This becomes important as most of these algorithms are stochastic and therefore should be investigated statistically. Hörl et al. (2020) is currently following this path of analysis.

For the specific case of French cities, many potential improvements are possible. More detailed OD flows by commuting mode could be used; the assignment of income could be improved by more elaborate models; destination data could be enriched by attractivity levels. Yet, we believe that the current state of our synthetic population of Île-de-France already provides value, not only for transport simulation, but also for more aggregate statistical analysis on mobility, infrastructure assessment and many other fields. As our code will be made available as open source before hEART 2020, we want to use the occasion to invite researchers to have a look at the framework, make use of the output and collaborate in open-data, open-source transport research.

References

Kickhofer, B., Hosse, D., Turnera, K., and Tirachini, A. (2016) Creating an open MATSim scenario from open data: The case of Santiago de Chile.

Ziemke, D., Kaddoura, I., and Nagel, K. (2019) The MATSim Open Berlin Scenario: A multimodal agent-based transport simulation scenario based on synthetic demand modeling and Open Data.

Horni, A., K. Nagel and K. W. Axhausen (2016) The multi-agent transport simulation MATSim, Ubiquity Press, London.

Lopez, P. A., Behrisch, M., Bieker-Walz, L., Erdmann, J., Flötteröd, Y., Hilbrich, R., Lücken, L., Rummel, J., Wagner, P. and E. Wiessner (2018) Microscopic traffic simulation using SUMO, *21st International Conference on Intelligent Transportation Systems (ITSC)*, Maui, HI, pp. 2575-2582.

Auld, J., M. Hope, H. Ley, V. Sokolov, B. Xu and K. Zhang (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.

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

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

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

Hörl, S., and Axhausen, K. W. (2020). Relaxation-discretization algorithm for spatially constrained secondary location assignment, 99th Annual Meeting of the Transportation Research Board.

Hörl, S., Balac, M., and Axhausen, K.W. (2020) Reproducible large-scale agent-based transport scenarios: A case study for Île-de-France, In preparation.

Hörl, S., Balac, M., and Axhausen, K. W. (2019, June). Dynamic demand estimation for an AMoD system in Paris. In 2019 IEEE Intelligent Vehicles Symposium (IV) (pp. 260-266). IEEE.

Sallard A., Hörl, S. and Balac, M. (2020) Agent-based scenario of Sao Paulo Metropolitan Area, In preparation.

Balac M. (2020) Agent-based scenarios of Los Angeles and San Francisco, In preparation.

MARG (2016) PopGen: Synthetic Population Generator [online]. Mobility Analytics Research Group. Available at: http://www.mobilityanalytics.org/popgen.html, Accessed 06.02.2020.