Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 257 (2025) 996–1001

Procedia Computer Science

www.elsevier.com/locate/procedia

The 14th International Workshop on Agent-based Mobility, Traffic and Transportation Models, Methodologies and Applications (ABMTrans 2025)

A first approach towards an automatized preparation of input data for the agent-based demand model TAPAS

Daniel Krajzewicz[*,1], Matthias Heinrichs[1], Antje von Schmidt[1], Alain Schengen[1], Simon Nieland[1]

[1]*Institute of Transport Research, German Aerospace Center, Rudower Chaussee 7, 12489 Berlin, Germany*

Abstract

In order to prepare data needed to simulate a new area using the agent-based demand model TAPAS a large variety of data has to be collected, consolidated, and converted. In the past, this included the purchase of commercial data about the activity places within the area to model, collection of available data about the socio-demographics of the area, and subsequent enrichment of both and the disaggregation of the population. Often, the process of preparing a new scenario could take few months. For allowing a more agile use of TAPAS, we develop a data import and preparation tool that allows preparing a TAPAS representation of an area using open data and needs only minor manual configuration. This paper outlines the methods used by the tool, presents initial results, and gives the next steps to be performed.

Keywords: agent-based demand model; data processing; open data.

[*] Corresponding author. Tel.: +49-30-67055-273; fax: +49-30-67055-283.
E-mail address: daniel.krajzewicz@dlr.de

10.1016/j.procs.2025.03.128

1. Introduction

The agent-based demand model TAPAS ("Travel Activity Pattern Simulation") [1][2] is a development of the German Aerospace Center (DLR), available as open source since the year 2020. TAPAS is a very fine-grained demand model that considers different socio-demographic attributes of the population for determining the persons' mobility behaviour. In addition, persons are grouped into households what allows to simulate the shared use of resources, e.g. available cars. During the data preparation process, each household is assigned to a dwelling within the simulated area. The high granularity of the representation and its high level of detail counts as well for the activity places within the modelled area. Every place that allows to perform certain activities at is represented by its geospatial location and capacity. In addition, activity places may be grouped to avoid using motorised modes of transport when, e.g., changing between shops within a bigger mall.

Due to the high resolution, building the representation of an area – usually a bigger city or parts of a federal state in Germany – is a time-consuming process. Information about the socio-demographics have to be collected from different sources, aligned against each other, disaggregated, and distributed over the dwellings within the area. Even though the last three steps were already automatized using the SYNTHESIZER tool [3], the collection and preparation of data had to be performed manually, as information about socio-demographics is usually available in Germany, yet often from different sources and often for spatial areas of different size. Data about the area's activity places were usually bought from external suppliers and had to be extended by capacities afterwards.

This paper presents an approach for generating TAPAS descriptions faster and using data available for the complete area of Germany, with options to use existing, yet more coarse data to set up regions throughout Europe. The remainder is structured as following: First, the data types needed by TAPAS are outlined in Section 2. Then, the new data preparation process is described in Section 3. A discussion of this attempt's shortcomings and meaningful extensions is given in Section 4. The paper closes with a summary in Section 5.

2. Data needed by TAPAS

TAPAS needs information about a) the population, including both, single persons as well as households the persons are assigned to, b) the places within the area where different activities can be performed at, and c) matrices describing the performance of different transport modes between and within traffic assignment zones. The information is stored in a Postgres database with the PostGIS extension installed. It is described in a greater detail in the following.

Albeit both, activity places, and the population are stored and used in a disaggregated manner – for each location and each household individually – both, locations and dwellings of households are assigned to traffic analysis zones (TAZ). The TAZ describe areas with a homogeneous mobility behaviour and are usually obtained from local administrations. The major need for using TAZs within TAPAS is to reduce the number of origin/destination relationships that have to be covered by the transport performance matrices.

The population consists of single persons with the following attributes: age, sex, employment status (not working, working full time, working half times), public transport ticket availability, driver licence availability, education level, and budgets for using an own motorised vehicle and for using public transport. The persons are grouped into households, which are defined by additional data about available cars, income, and existence of children in the household. When computing the mobility within an area, TAPAS iterates over the households, first, then over the persons that belong to the household. The assignment of the vehicles available over the day to one of the household's persons is described in [4].

The locations are described using their position and an activity code, together with the information about the TAZ they are located in. The activity codes are variable (see also [5]), making an additional mapping between the locations' activity codes and TAPAS activities necessary.

TAPAS' mode choice model uses matrices that describe the performance of the major transport modes – walking, bicycling, public transport (PT) and motorised individual transport (MIT) – between the TAZ. The performance is described in means of distances (in meters), travel times (in seconds), as well as access and egress times (both in seconds). In addition, a further matrix holds the number of interchanges between two cells for PT.

3. Proposed data preparation process

The new methodology for a transferable generation of TAPAS inputs is given in the following. A data set for Berlin in the year 2017 generated using the initial approach is put against the results of the new methodology. The sub-topics of the zoning, the computation of the population, of the activity locations, and of performance matrices are discussed separately.

3.1. TAZ

A main assumption for simplifying the preparation process was to use a rectangular grid for TAZ. This should ease the assignment of dwellings, households and activity places to TAZ, especially when moving to the replication of European areas apart from Germany, where population data is available aggregated in an INSPIRE-conform grid of 1 km × 1 km. As such, in a first step, the TAZ grid is built. The only information needed to achieve it is the boundary of the area to replicate. Given the TAZ, the population is built, first, then the locations.

3.2. Population

We currently use the address dataset from the Bundesamt für Kartographie und Geodäsie (BKG) which contains the addresses of buildings with the attached information about the number of inhabitants and households. Given this information, we can build the population for each cell by iterating over the dwellings within this dataset and generate the according number of persons and households. This process involves external data, namely the distribution of the population by age and sex, and the distribution of working persons by age and sex. Internally, the population age statistics are split into persons below and above 20 years, obtaining “children” and “adults”. In addition, the share of women within the population is computed. For each household in the currently regarded dwelling, a single person is generated, first. Then, persons are added to the households randomly.

When generating a person, the first person of a household is always assigned to be an adult person. Its age is chosen from the age distribution of adults, the sex is selected based on the previously computed share of female persons in the population. The second person may be a second adult or a child. All subsequently added persons are children. If the second person is an adult, the age is computed by adding randomly ±5 years to the first persons’ age and the opposite sex of the first person is chosen. The age of all adults is normalised to be between 16 and 99 years. The age of children is selected from the child age distributions and the sex is again chosen randomly, considering the share of female persons in the population.

The employment status of adults is selected from the external employment distribution, considering the person’s sex. Initially, all children older than 6 years are assigned to be pupils, first. Then, children above 15 years are assigned randomly to be visiting a professional training, children above 17 years are assigned randomly to be visiting a professional training or being a student.

The obtained population for the city of Berlin, representing the status in 2024 is given in Figure 1b and is put against the initially generated representation of the year 2017 given in Figure 1a. The reasons for the higher fluctuations between the age groups in Figure 1a have yet to be determined. Figure 1b reveals that the information about employment is given in steps of five years.

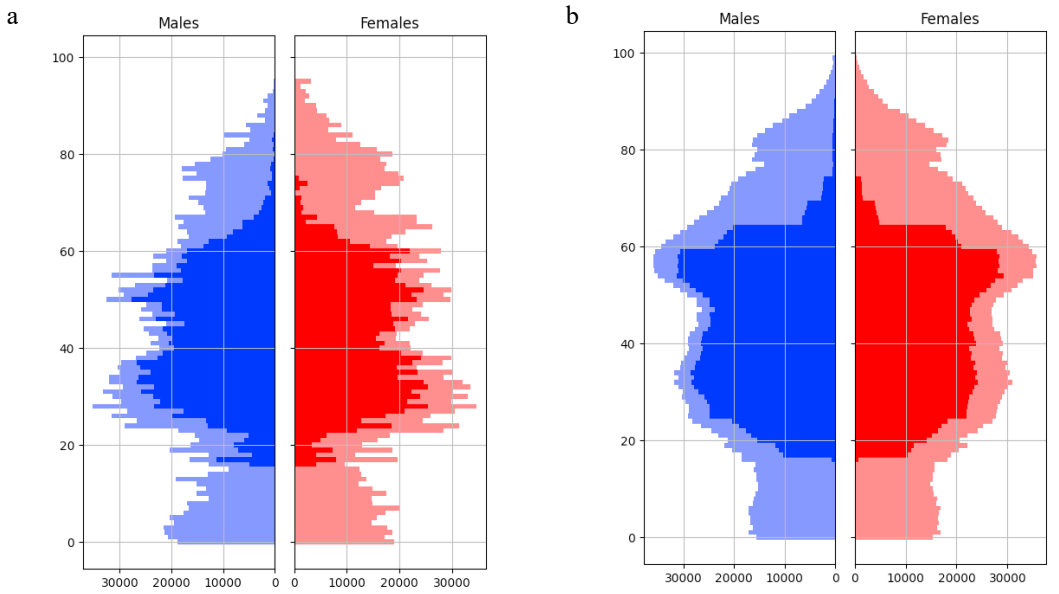

Fig. 1. Comparison of the age and employment distributions by sex between settings generated using the old (a) and the new (b) method; please note an offset of about 7 years (2017 for a, 2024 for b) between the data sets.

As soon as all persons have been generated for a household, the household's income is computed, assuming a Gaussian income distribution. In addition, the cars belonging to the household are computed, using car ownership share which is a further input parameter to the method.

3.3. Activity Locations

Activity locations are extracted from OSM. For this purpose, a new scheme that maps all *building*, *shop*, *amenity*, *healthcare*, *tourism*, *leisure*, *sport*, and *office* keys with their value sets to location codes has been developed. This mapping allows to assign a certain activity code to each key/value pair, which maps to a set of different activities. E.g., a school is assigned to the code "1002001001" which represents the activities "education-primary school" as well as "education-further education", partially performed at schools. In addition, most of the activity places are as well used as working locations.

The location instances are extracted from an OSM data using a script from the UrMoAC [6][7] package, an open source tool form computing accessibility measures. The method described herein uses the so obtained instances, yet re-interprets the original OSM dataset to obtain the capacity of the location. If given, the capacity information is used directly. Otherwise, the building's footprint and its number of levels are used, together with Bosserhof factors, to determine the capacity for the visitors of the infrastructure as well as the number of persons working at the infrastructure. A grouping of locations is currently not implemented.

Figure 2a shows a comparison of the numbers of locations different activities can be performed at between the old TAPAS input data set, and the one generated using the method described herein. It should be noted that the original (old) data set used commercial data, what made the generation of the simulation expensive. Given that both use completely different input data, the match between the numbers is satisfying. Figure 2b shows the according capacities. Here, a higher mismatch is visible, making further investigations necessary.

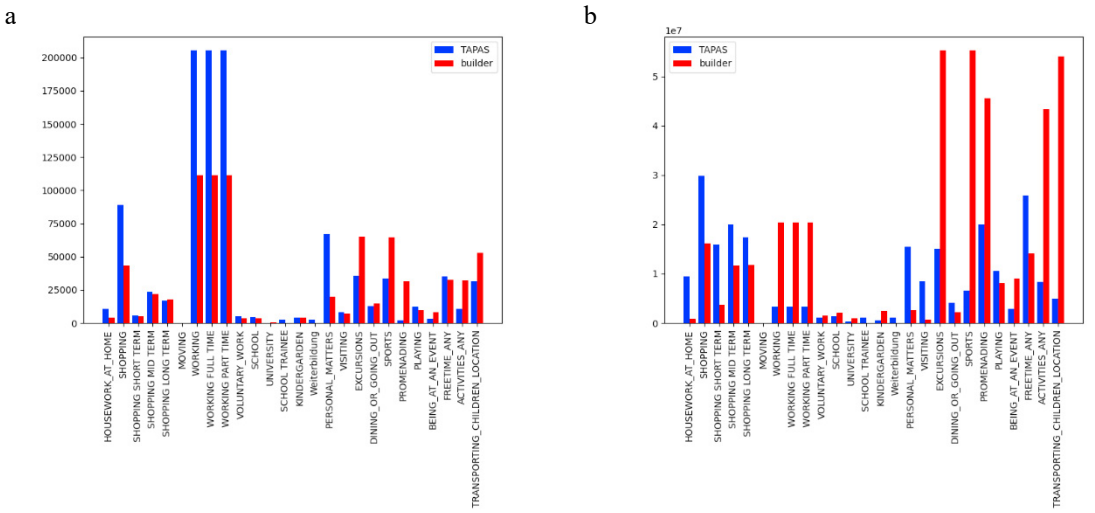

Fig. 2. A comparison of locations within the representation of Berlin between the old and the new method per activity type in terms of (a) their numbers, and (b) their capacities.

3.4. Performance Matrices

The travel time, distance, access, and egress matrices for trips between two cells are computed using the UrMoAC tool as described in [8]. Currently, this is the most time-consuming process of the preparation of an area and is thereby set up as an own workflow, yet using the previously generated TAZ and buildings data. Three variants are supported: "fast" which routes between the TAZ centres only, "representatives" which randomly selects a predefined number of origin / destination locations per TAZ, and "complete" which routes between all buildings within each TAZ.
The computation of inner-TAZ performance measures (see [9]) is not yet included in the work flow.

4. Current Shortcomings and Planned Extensions

The first tests showed several shortcomings which will be discussed in the following. Albeit using a rectangular grid eases the mapping of given data, some cases occur where a different spatial aggregation makes sense. This is, e.g. the case when a TAZ is divided into parts by a river or any other natural obstacle. This is a well-known topic in transport modelling and the proposed system should be extended by according methods for splitting and re-joining TAZs.
Regarding the population, the random selection of person attributes from statistics is assumed to be error-prone and neglects inter-dependencies between attributes. As soon as further statistical data will be collected, the system should therefore be extended by methods for aligning different statistics to compute a valid population. Methods such as IPF/IPU could be employed here, what is already part of the initial preparation process performed by the SYNTHESIZER. In conjunction, the system should be extended by methods for extrapolating the population for modelling future scenarios. As well, it seems meaningful to split students from the main population, due to their specific age range and their habits in terms of dwelling selection.
The generation of households currently disregards the distribution of household sizes, which is available on regional scale. As well, the current household computation method generates very uniform households, which neither resemble same-sex households nor children with an age above 20 years living with their parents.
The differences in the capacities show that the individual per-activity capacities should be viewed in a decomposed way and a weighting of each component should be added to the computation. E.g., places like forests have a high capacity due to their large area, yet only the roads within them are really accessible and used. In contrary, locations at which personal matter activities are performed seem to be under-represented in OSM (see [10]) and should be thereby upscaled.

Finally, missing methods for grouping activity locations and for computing inner-TAZ performance measures must be added to the workflow.

5. Summary

The method proposed herein offers the possibility to set up new areas for TAPAS fast and at low cost while reducing the need to collect necessary data by using data available throughout Germany. Yet, different shortcomings to the original process are known and should be diminished in subsequent development steps. In addition, data sources for generating areas beyond Germany need to be collected and their quality – especially when using crowd sourced data as in the case of OSM – needs to be benchmarked.

Acknowledgements

Gefördert durch:

Bundesministerium für Wirtschaft und Klimaschutz

aufgrund eines Beschlusses des Deutschen Bundestages

The work presented herein was performed in the scope of the project “Interoperables Management für Bidirektionales Laden für den optimierten, resilienten Stromnetzbetrieb mit innovativen Geschäftsmodellen” (InterBDL) co-funded by Germany’s Federal Ministry for Economic Affairs and Climate Action under the identifier 01MV23025.

References

[1] Heinrichs, Matthias, Daniel Krajzewicz, Rita Cyganski, and Antje von Schmidt (2016) “Disaggregated car fleets in microscopic travel demand modelling”, *The 7th International Conference on Ambient Systems, Networks and Technologies* (ANT 2016), DOI: 10.1016/j.procs.2016.04.111.

[2] DLR Institute of Transport Research, (2021), https://github.com/DLR-VF/TAPAS; last visited on 16.12.2024.

[3] von Schmidt, Antje und Cyganski, Rita und Krajzewicz, Daniel (2017) Generierung synthetischer Bevölkerungen für Verkehrsnachfragemodelle - Ein Methodenvergleich am Beispiel von Berlin. In: HEUREKA'17 - Optimierung in Verkehr und Transport, Seiten 193-210. FGSV-Verlag. HEUREKA'17, 2017-03-22 - 2017-03-23, Stuttgart, Deutschland. ISBN 978-3-86446-177-4.

[4] Beige, Sigrun und Heinrichs, Matthias und Krajzewicz, Daniel und Cyganski, Rita (2017) Who gets the key first? Car allocation in activity-based modelling. International Journal of Urban Sciences, Seiten 1-15. Taylor & Francis. doi: 10.1080/12265934.2017.1351389. ISSN 1226-5934.

[5] Radke, Andreas und Heinrichs, Matthias (2021) Using Probability Distributions for Projecting Changes in Travel Behavior. Sustainability, 13 (18), Seite 10101. Multidisciplinary Digital Publishing Institute (MDPI). doi: 10.3390/su131810101. ISSN 2071-1050.

[6] Daniel Krajzewicz, Dirk Heinrichs and Rita Cyganski (2017) Intermodal Contour Accessibility Measures Computation Using the 'UrMo Accessibility Computer'. International Journal On Advances in Systems and Measurements, 10 (3&4), Seiten 111-123. IARIA.

[7] DLR Institute of Transport Research, (2024), UrMoAC-0.8.2, zenodo, doi: 10.5281/zenodo.13234444.

[8] Krajzewicz, Daniel und Schengen, Alain (2024) Computation of mode-dependent travel time matrices for an agent-based demand model computed using a standalone accessibility tool. Procedia Computer Science (238), Seiten 779-784. Elsevier. doi: 10.1016/j.procs.2024.06.091. ISSN 1877-0509.

[9] Heinrichs, Matthias und Cyganski, Rita und Krajzewicz, Daniel (2021) Address-based computation of intra-cell distances for travel demand models. In: 12th International Conference on Ambient Systems, Networks and Technologies, ANT, 184, Seiten 123-130. Elsevier. The 12th International Conference on Ambient Systems, Networks and Technologies ANT2021, 2021-03-23 - 2021-03-26, Warsaw, Poland. doi: 10.1016/j.procs.2021.03.023. ISSN 1877-0509.

[10] Friedrich, Til (2024) Entwicklung eines Diversitätsindikators zur Planung nahräumlicher Versorgungsinfrastrukturen am Beispiel Berlin-Brandenburg. Bachelorarbeit, Technische Hochschule Wildau.