Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 257 (2025) 974–979

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)
April 22-24, 2025, Patras, Greece

Modelling the effects of anonymization in O/D matrices obtained from mobile phone data

Daniel Krajzewicz[*,1], Hans Arne Trukenbrod[2], Martin Haerst[3], Louis Touko Tcheumadjeu[4]

[1]*Institute of Transport Research, German Aerospace Center, Rudower Chaussee 7, 12489 Berlin, Germany*
[2]*T-Systems International GmbH, Winterfeldtstraße 21, 10781 Berlin, Germany*
[3]*Stadt Köln – Die Oberbürgermeisterin, Amt für Verkehrsmanagement, Willy-Brandt-Platz 2, 50679 Köln, Germany*
[4]*Institute of Transportation Systems, German Aerospace Center, Rutherfordstr. 2, 12489 Berlin, Germany*

Abstract

Mobile phone data promise a ubiquitous traffic surveillance without the need of additional hardware. Yet, within most European countries, the data have to be anonymized to disallow the recognition of individuals. To understand the effects of applying the anonymization methodology, we use the results of an agent-based demand model as the ground truth and apply a well-described anonymization method to the trips performed by the simulated population. Besides investigating the segmentation of the area and the time aggregations used by the data supplier T-Systems, we iterate over different cell sizes and aggregation times. The results show how the completeness of the reported data deteriorates with a decrease of cell size and time span used for aggregation. As well, the original segmentation of the area using a variable grid shows to be a good compromise between a fine-grained resolution and the reported number of trips.

Keywords: mobile phone data; modelling; anonymization effects.

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

10.1016/j.procs.2025.03.125

1. Introduction

Within the German project “MoCKiii” [1], O/D matrices obtained from anonymized mobile phone data are used to support an algorithm that estimates the modal split within the city of Cologne with the number of rides within the city. These matrices do not include all rides as not all persons use a phone and because of the necessary anonymization process. For using this data for calibration, it is thereby necessary to know how complete the matrices are and what artefacts the anonymization process may introduce. Within the study presented herein, we use the agent-based demand model TAPAS [2][3] as the ground truth for a region’s mobility and apply the anonymization rules to the model’s results to investigate these topics.
The remainder is structured as following: We describe the used agent-based demand model TAPAS shortly in Section 2. The anonymization rules as well as the process of replicating them are given in Section 3. The results are presented and discussed in Section 4. We close with a short summary in Section 5.

2. The agent-based demand model TAPAS

TAPAS is an agent-based demand model where every inhabitant of the modelled area is individually represented using her/his socio-demographic attributes like age, gender or employment status and information about available mobility options, such as a bike, a driving license or a public transport subscription. In addition, persons are grouped into households what allows for simulating the usage of shared mobility resources, such as a family car. Both, the persons as well as the households are generated from diverse administrative data that are disaggregated using the IPU method by a supporting tool named SYNTHESIZER [4] during the preparation of the simulation input data. Besides a representation of the population, TAPAS needs information about the transport network, including the public transport, in form of travel time and distance matrices, see also [5]. In addition, information about the locations of activity places, like schools, work places, shops, etc. within the modelled area are needed.
Given this input, TAPAS iterates over the households and persons. For each person, a daily activity plan that matches the person’s socio-demographics is chosen [6] from a list of about 50,000 plans obtained from the German mobility survey “Mobilität in Deutschland” (MiD) [7]. For each of the activities stored in the plan, TAPAS determines the according location within the modelled region as well as the mode of transport used by the person to access it. Both, the distribution of the travelled distances as well as the mode choice model are again derived from the MiD.
The result of a TAPAS run consists of a list of trips performed during a usual working day for each of the simulated persons. Each trip is described by the positions it starts and ends at as well as its starting time and duration, and the activity performed after the trip, among other. Figure 1 depicts the workflow of TAPAS.

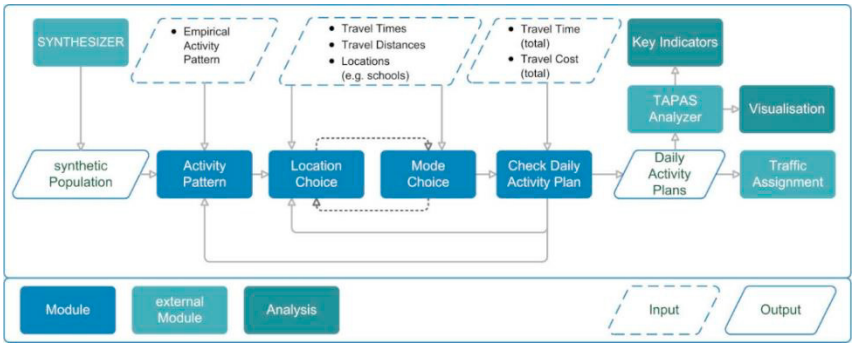

Fig. 1. TAPAS workflow.

Given the amount of needed input data, currently only few German regions are modelled in TAPAS. The results presented herein use the replication of Brunswick. This region has been chosen because it covers rural areas including commuting traffic.

3. Anonymization of mobile phone data

The mobile phone data used in the MoCKiii project are supported by Motion Data[2] of T-Systems in form of O/D matrices. The referred cells match the orientation and alignment of the INSPIRE grid for Germany, yet the cells have varying sizes, ranging from 0.25 km^2 to 64 km^2. The matrices are supplied in two different time aggregation spans, one for each hour of the day and one aggregated over the complete day. The matrices are submitted as plain csv-files, where each line contains the referenced time, the IDs of the source and the destination cell, and the number of trips performed between these cells that started within the given time span.

The anonymization is performed by T-Systems as described in the following. First, the number of seen trips that started within the reported time span between two cells is counted. This value is then upscaled using the inverse of the regional share of Telekom users. If the result falls below a threshold k, being currently $k=30$, the O/D pair is dismissed, otherwise it is reported.

We apply this anonymization procedure on the results obtained from TAPAS. Besides using the original spatial grid ("orig.") as used by T-Systems and time aggregation spans of 1 h and 24 h, we additionally iterate over different cell sizes – with side lengths of 500 m, 1000 m, 2000 m, 4000 m, and 8000 m – and over time aggregation spans of 1 h, 2 h, 4 h, 6 h, 12 h, and 24 h. We start with T^{all} being all trips reported by TAPAS. T-Systems uses a threshold of 12 minutes to recognize if a trip ends. Short shopping activities are thereby not included within the O/D matrices. This differs from TAPAS which also reports trips with a shorter stay time at the destination location. In the research presented herein, a TAPAS trip is joined with his successor if the stay time falls below the threshold of 10 minutes, resulting in a set of trips named T^{base}. On the other hand, TAPAS does e.g. not report waiting times at a public stop halt, what may be recognized by T-Systems as a stop. It is hardly possible to develop a balanced benchmarking for this sub-problem, which is as well not the major topic of this publication. Therefore, T^{base} is used as the base set of all trips to compare the effects of the anonymization against in the following.

When processing the data, trips that start before the day begins or after the day ends are removed. The persons are then stochastically assigned to be a Telekom user or not based on the assumed German mobile phone market share of 30%. The trips from T^{base} performed by users visible to Telekom are called T^{seen}. They are upscaled by being multiplied by the inverse Telekom share resulting in $T^{upscaled}$. T-Systems states that persons with an age < 6 years or ≥ 90 years usually do not have a mobile phone and are thus not tracked. Though, we do not remove these persons from our sets as a) children younger than six years old are not simulated by TAPAS and b) this would reduce the comparability of the processing steps.

Next, for each trip, the origin and destination cells are determined using the geo-coordinates of a trip's starting and ending location, respectively. The number of the so obtained O/D relations is then aggregated within the respectively computed time interval, using the trip's starting time for determining the interval itself. After processing all trips within a time interval, all included O/D relations are examined for being reported or not. If the number of trips of an O/D relationship is below k=30, all trips in the O/D relation are assigned to T^{too_low} meaning that they were anonymized and will not be reported, otherwise they are assigned to T^{kept}. Figure 2 shows the anonymization and de-anonymization processing steps using an example. Please note that the two rightmost figures hide the information in which cases the upscaling yield higher numbers than the size of T^{base}.

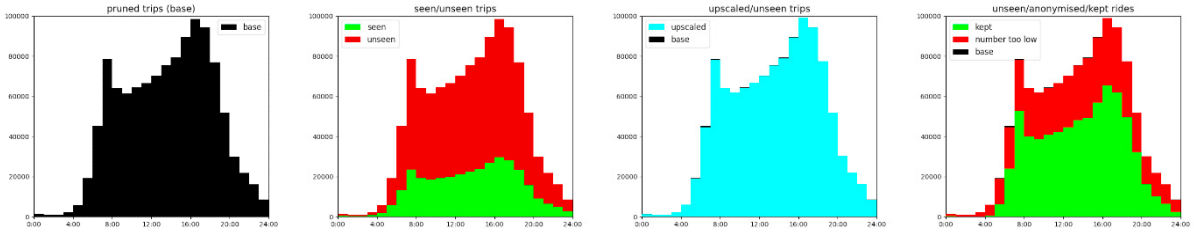

Fig. 2. Process steps outline based on Brunswick data and using a time aggregation of 1 h and a spatial aggregation into cells with a side size of 2000 m as example.

[2] https://dih.telekom.com/en/motion-data

Two further classifications were implemented: a) trips in T^{seen} are assigned to T^{same_cell} if the origin and the destination locations fall into the same cell, otherwise to T^{od}, b) the trips in both, T^{seen} and T^{too_low}, are additionally distinguished by the used mode of transport.

4. Results

4.1. Dependence on Spatial and Time Aggregation

Both, spatial aggregation as well as aggregation over time have an effect on the number of dismissed O/D-pairs. With a growing spatial aggregation, the probability of trips to be assigned to the same O/D-pair increases. As well, with a growing duration of the aggregation time span, more trips are found within the O/D pairs. Both affects the number of trips assigned to T^{too_low}. Figure 3a shows the (upscaled) shares of trips that are dismissed due to a too low number of persons (T^{too_low}/T^{base}). The shares of the lost O/D-connections are shown in Figure 3b.

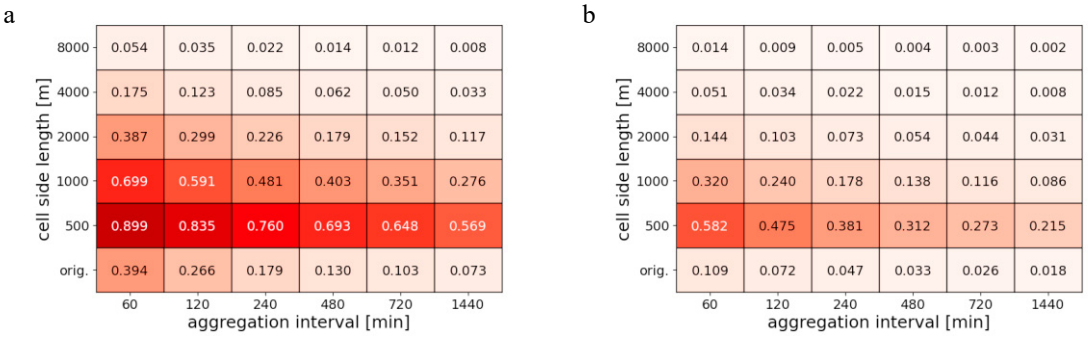

Fig. 3. (a) shares of trips dismissed due to anonymization; (b) shares of not reported o/d connections; both in dependence to the aggregation over time and spatial aggregation.

It is interesting to observe that the number of reported trips (T^{kept}) that take place within a cell (T^{same_cell}) has a different development than the number of trips between two different cells (T^{od}), see Figure 4. Of course, with an increase in the cell size, more trips take place within the cells. The trips in T^{same_cell} are less sensitive to the aggregation time span in comparison to trips between two different cells.

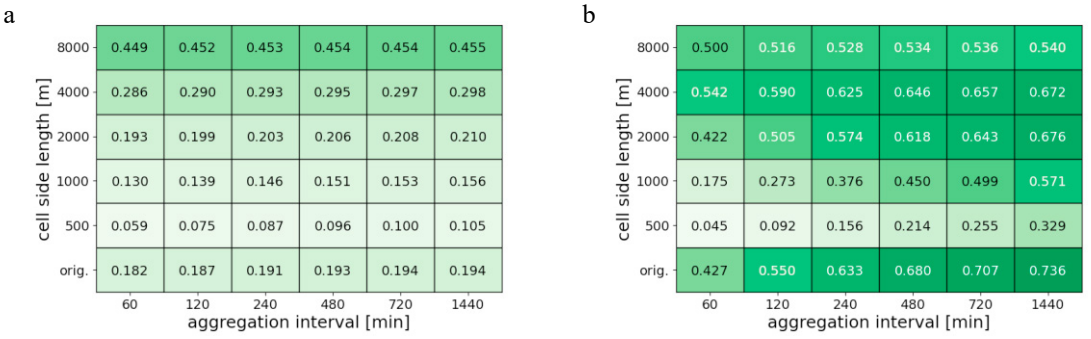

Fig. 4. (a) shares of reported (non-anonymized) trips within cells; (b) shares of reported (non-anonymized) trips between two different cells; both in dependence to the aggregation over time and spatial aggregation.

4.2. Mode-Dependent Effects

Figure 5 shows the shares of reported trips (T^{kept}) in dependence to the used transport mode for the investigated time and spatial aggregations. It shows that trips performed by foot are removed least often, followed by bike, car, and

public transport in this order. This can be explained by the differences in the travelled distances using the respective mode yielding in a larger spread across destination cells.

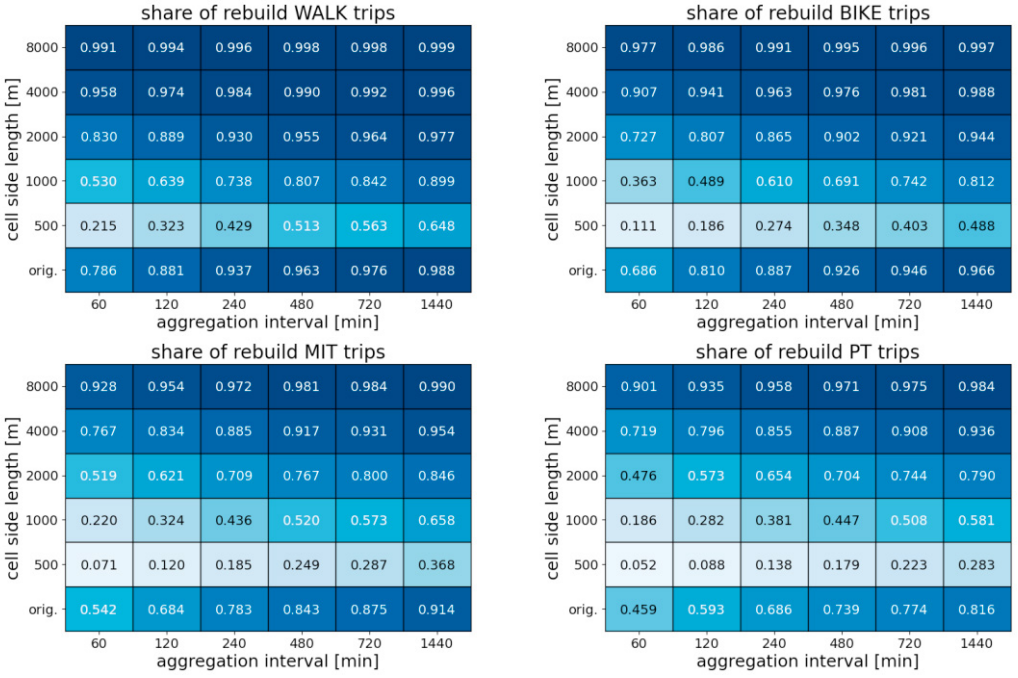

Fig. 5. Shares of reported trips (T^{kept}) in dependence of the used mode of transport.

4.3. Qualitative Differences

As already visible in Figure 2, the rebuild curve differs not only in numbers, but as well in shape. Figure 6a shows the development of the curves for T^{base}, T^{kept}, and T^{too_low}. The normalized deviations of T^{kept} and T^{too_low} from T^{base} are shown in Figure 6b. In this example, the mobility pattern at night, between 1:00 and 5:00 are too sparse and are anonymized. This has to be handled with care – TAPAS diaries end at the end of the day, and persons that remain moving may be missing in the early hours. In addition, Brunswick, which was used as test case has only a quarter of Cologne's population, where more people will be mobile.

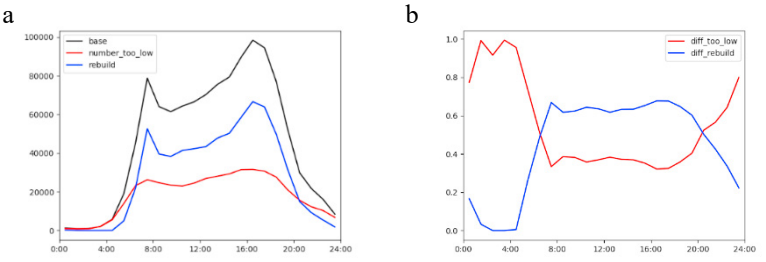

Fig. 6. (a) absolute numbers of all (T^{base}), reported (T^{kept}) and missing (T^{too_low}) trips; (b) shares of reported (T^{kept}) and missing (T^{too_low}) trips.

5. Summary

To investigate the effects of anonymization on the quality of reported trips from mobile phone data, the results of the agent-based demand model TAPAS were used as a ground truth. Starting with the complete number of trips as reported by TAPAS, the anonymization algorithm was applied, iterating over different spatial and time aggregations.
The results show how the completeness of the reported data depends on the aggregations, what, in conjunction, affects the shares of reported trips per used mode of transport. As well, the original segmentation of the area using a variable grid proofs to be a good compromise between a fine-grained resolution and the reported number of trips.
The study presented herein covers only one of the areas replicated in TAPAS. In subsequent steps, possible differences between different areas or cities should be examined. As well, the methods should be evaluated using a demand model that describes the mobility of the city of Cologne – either employing a TAPAS model of the city or the city's transport model. In addition, the effects of the used method for recognizing whether a trip has ended have not been discussed herein and should be a matter of a future investigation. The anonymization factor, currently being k=30 is meant to be changed to k=5 in the future. Future investigations should resemble this change.
Albeit not being based on real data but on modelling results with own artefacts and shortcomings, we assume that our results contribute to a deeper understanding about how anonymization and the fact that only a subset of the population is seen affect the quality of mobile phone data for monitoring mobility.

Acknowledgements

Gefördert durch:

Bundesministerium für Digitales und Verkehr

aufgrund eines Beschlusses des Deutschen Bundestages

The work presented herein was performed in the project "Mobilitäts-Cockpit Köln - innovativ - integrativ - intelligent" (MoCKiii) co-funded by Germany's Federal Ministry for Digital and Transport under the identifier 16DKVM007B.

References

[1] City of Cologne. (2022) "Mobilitäts Cockpit Köln – MoCKiii." (German), web page, last visited on 7th of October 2023.

[2] 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.

[3] Heinrichs, Matthias, Daniel Krajzewicz, Rita Cyganski, and Antje von Schmidt (2017) "Introduction of car sharing into existing car fleets in microscopic travel demand modelling", *Personal and Ubiquitous Computing*, pp. 1–11, Springer, ISSN 1617-4909, DOI: 10.1007/s00779-017-1031-3.

[4] von Schmidt, Antje, Rita Cyganski, and Daniel Krajzewicz (2017) "Generierung synthetischer Bevölkerungen für Verkehrsnachfragemodelle - Ein Methodenvergleich am Beispiel von Berlin" *HEUREKA'17 - Optimierung in Verkehr und Transport*, pp. 193-210. FGSV-Verlag. ISBN 978-3-86446-177-4.

[5] 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), pp. 779-784. Elsevier. doi: 10.1016/j.procs.2024.06.091. ISSN 1877-0509.

[6] Radke, Andreas, and Matthias Heinrichs (2021) "Using Probability Distributions for Projecting Changes in Travel Behavior." *Sustainability*, 13 (18), pp. 10101. doi: 10.3390/su131810101. ISSN 2071-1050.

[7] Gruschwitz, Dana, Johannes Eggs, Claudia Nobis, and Angelika Schulz (2018) "Analyses of mixed survey mode effects in the German national travel survey 2017." *Transportation Research Procedia*, 32, pp. 339-350. Elsevier. doi: 10.1016/j.trpro.2018.10.062. ISSN 2352-1457.