Available online at www.sciencedirect.com

ScienceDirect

Transportation Research Procedia 97 (2026) 542–552

Transportation Research Procedia

www.elsevier.com/locate/procedia

13th International Conference on Transport Survey Methods

A framework for using mobile phone data in population synthesis for agent-based modeling

Ann-Sophie Voss[a,*], Pia Tulodetzki[a], Tim Wörle[a], Martin Kagerbauer[a], Peter Vortisch[a]

[a]*Institute for Transport Studies, Karlsruhe Institute of Technology, Otto-Ammann-Platz 9, 76131 Karlsruhe, Germany*

Abstract

In this study, we introduce a methodology that follows the conventional approach of generating a synthetic population (SynPop), which is subsequently enhanced and refined by geographical information using mobile phone data (MPD). The overarching goal is to enable a more specific trip distribution based on the established 4-stage model of transport demand modeling in our study area the city of Darmstadt and its surrounding rural areas in Germany. By employing a mixed exact-probabilistic data-matching approach, the methodology enhances representativeness and granularity, addressing biases and limitations of current datasets. Overall, six key attributes with a mixed approach of exact and non-exact matching were utilized for the matching process. The matching was executed iteratively using a greedy algorithm, continuing until no data points remained. A comparison of the two applied distance metrics, Nearest Neighbor matching using Propensity Scores and Mahalanobis distance each with and without a caliper, revealed, after validating the balance, that the Mahalanobis distance without a caliper achieved more reliable matching for our data. Nevertheless, the data basis must be more precisely aligned with the matching parameters in the future.

Keywords: Travel Demand Modeling; Data Matching; Mobile Phone Data

1. Introduction

People travel to various places to carry out their individual activities throughout the day. When examining these mobility patterns at the population level, a complex network of mobility structures emerges. A thorough understanding of these structures is crucial for developing travel demand models, which may serve as the foundation for implementing targeted mobility management measures (Demissie et al., 2019). To do so, it requires data that represents such mobility patterns. Traditionally, household surveys have been used for this purpose, as they can track individual mobility decisions and their geographical locations over an extended period of time (Stopher and Greaves, 2007). As this approach is prone to low response rates, inaccuracy in reporting, thus incomplete and outdated data, and an expensive implementation leading to small sample sizes, it is questionable whether this method involves a distorted representation of mobility patterns (Wilson, 2004; Stopher et al., 2007; Stopher and Greaves, 2007).

* Ann-Sophie Voss. Tel.: +49 721 608-46002
E-mail address: ann-sophie.voss@kit.edu

2352-1465
10.1016/j.trpro.2026.04.192

In recent years, the use of anonymized mobile phone data has become an emerging option. This readily-available and cost-effective data is made accessible by mobile network providers (Wang and Chen, 2018). The benefit lies in the geographical data's capability to offer detailed insights into individuals' trajectories, with minute-level precision, spanning extended periods. Given the widespread use of smart devices nowadays, large-scale samples can be obtained on an aggregated level (Wang and Chen, 2018; Calabrese et al., 2013). However, there are several drawbacks to consider, which preclude the exclusive use of mobile phone data: the lack of connection to causal factors in decision-making, the over-representation of demographic groups with more frequent mobile phone ownership, and the bias towards phone-usage in certain locations (Smoreda et al., 2013; Li et al., 2024). To balance out these trade-offs, it appears to be promising to explore the combination of both approaches when modeling mobility patterns on an agent-based level.

In this paper, we introduce a methodology that follows the conventional approach of generating a synthetic population (SynPop) at the individual and household levels, which is subsequently enhanced and refined using mobile phone data (MPD) for the city of Darmstadt, Germany, and its surrounding areas. The overarching goal is to enable a more specific trip distribution based on the established 4-stage model of travel demand modeling. As part of this process, the generated activity plans are enriched by assigning destinations derived from mobile phone data, ensuring a higher spatiotemporal resolution of the trips. Aggregated metrics are employed to execute the matching process. A more detailed exposition of this approach will be provided in the following section.

1.1. Data matching approaches

Data-matching processes operate through a series of structured steps. A comprehensive overview is presented in the book by Peter Christen, which also serves as the basis for this paper (Christen, 2012). After data pre-processing, during which data is standardized and unified, the process moves to indexing. At this stage, records that are highly unlikely to form a data pair are filtered out, significantly reducing the computational effort required to compare every record in two databases. This step not only streamlines the matching process but also results in cost and time savings. Following indexing, data matching takes place, identifying data pairs. These pairs are then classified into either a match, a non-match or a potential match which, in case of the latter, needs to be evaluated manually. Lastly, matches are assessed in terms of their matching quality (Christen, 2012).

Within the data-matching step, different methods are employed to capture the wide range of matching approaches. These methods are described in detail in different sources such as those by Peter Christen and Elizabeth Stuart (Christen, 2012; Stuart, 2010). Overall, we divided into exact and non-exact matching approaches. Exact matching relies on predefined rules that operate on an if-then logic. This method is deterministic, avoiding the use of probabilities and uncertainties, and ensures a straightforward matching process (Hakak et al., 2019). However, a significant drawback of this approach is its inability to accommodate minor inconsistencies, such as typographical errors or transposed digits, which can lead to the exclusion of valid potential matches (Stuart, 2010).

Non-exact matching or probabilistic matching seeks to address this limitation by employing probabilistic models to determine whether two data points represent a match (Stuart, 2010). Instead of rigidly categorizing pairs as matches or non-matches, this approach calculates the probability of a match, allowing for a more nuanced evaluation. Probabilities are often derived using similarity measures, which are tailored to the type of data being compared, whether string-based or numerical. A similarity function is used to compute a similarity value, which ultimately provides an assessment of the similarity between two data points (Nagels et al., 2019). It is important to emphasize that this research does not focus on foundational studies of data matching or similarity measures. Instead, it evaluates the applicability of established methods for matching data. In our case, we employ a combination of exact and non-exact matching to maximize the generation of potential matches.

2. Methodology

2.1. Data

First, the two foundational data sets —SynPop and MPD— were prepared, serving as the basis for the subsequent data matching process.

2.1.1. Synthetic population

We generated a synthetic population (SynPop) by integrating household data on travel behavior with census data from the Hessian Office for Statistics (2022). This merging provides additional sociodemographic information, such as age and gender, in various geographical zones in Darmstadt and surrounding rural areas. The household data used for this study stems from *Mobility in Germany*, a cross-sectional survey conducted in 2017. It contains data from around 316,000 participants in about 156,000 households throughout Germany. The data set includes both sociodemographic information and a one-day travel diary (Nobis and Kuhnimhof, 2018). To adjust the data to our study area, it was filtered to include only urban households from Hesse and the surrounding federal states Baden-Württemberg and Rhineland-Palatinate. To do so, only households with completed questionnaires were included, resulting in a final data set of around 43,700 individuals from 20,500 households.

By combining census data and the household data using the Iterative Proportional Updating (IPU) algorithm, individual households are assigned and scaled to meet the population distribution of the census data, such as overall population size and demographic distributions within different geographical zones. This ensures that the synthetic population aligns with both the detailed survey data and the broader census-based demographic structure. Additionally, workplace information is incorporated using commuting matrices from the Office for statistics as well as workplace data. As part of the IPU, each individual is assigned a place of residence and, if they are employed, a place of work. The final data set of the synthetic population contains 819,442 individuals. After that, each individual is assigned a model-based activity plan created with actiTopp, a module integrated in the agent-based travel demand model mobiTopp (Hilgert et al., 2017). ActiTopp generates activities carried out during the course of a day including their order and respective length. Further information on mode or geographical locations are not considered at this point. As a result, each individual in SynPop is given an activity chain for one day comprised of the activity types Work (W), Home (H) and Other (O). Only trips to the workplace are counted as Work; all other purposes, such as trips to university or schools as well as business trips are counted as Other.

2.1.2. Mobile phone data (MPD)

The mobile phone data were obtained from *Invenium Data Insights GmbH*, a company specialized in processing such data. The data is generated when a phone connects to the nearest cell tower. As a result, anonymous information can be generated by assigning an ID to each individual phone and day. This data includes the start and end times of the trips in 5-minute intervals, the duration of the trips, and the origin and destination traffic cells. However, the spatial accuracy can vary: in urban areas, precision is approximately 100 meters, while in rural areas, it can be several hundred meters.

Once processed, the data set provides valuable insights into mobility patterns, such as the number of trips taken, the total distance traveled, and the spatial distribution of residential and work locations. The trip purposes are categorized into three predefined groups: Home (H), Work (W), and Other (O). In this specific case, the data set comprises data from Telefonica customers, a mobile network operator in Germany that covers approximately 25% of the German market. It includes two full weeks of data: from April 24 to April 30 and from May 22 to May 28, 2023, excluding weekends. These periods were chosen due to the absence of public holidays or other special events, such as school vacations. Geographically, the data is limited to the planning area around Darmstadt and the district of Offenbach.
The MPD has undergone various cleansing procedures to enhance the similarity of the initial data sets of SynPop and MPD. While the SynPop only includes the activity of individuals residing within the designated planning area, the MPD encompasses each individual who has remained in the planning area for a minimum duration exceeding the detection threshold (approximately 10 minutes). Consequently, individuals whose home location was outside the planning area were excluded, as well as those whose final journey occurred after midnight and those who completed only a single trip in a day. The final data set comprises about 1.5 million individuals on ten distinct working days.

To gain an initial understanding of SynPop and the MPD and to identify possible matching covariates, we looked at several descriptive statistics. Table 1 shows the ten activity chains that show the highest shares in the synthetic population and the corresponding share in the MPD. In both data sets, the Home-Other-Home chain constitutes the largest proportion. It is noticeable that some chains, such as HOHOH, only make up a small part in the MPD, while they occur frequently in the SynPop. In contrast, certain activity chains in SynPop accounts for less than 1% in the MPD. The disparity in proportions can possibly be attributed to the difference in the number of route chains between

the MPD and the SynPop, with the former containing about 4500 activity chains and the latter about 3300.

Table 1. Activity Chain Comparison between SynPop and MPD (chains with the largest shares)

	HOH	HOHOH	HWH	HWHOH	HOOH	HOHOHOH	HWOH	HOOOH	HOOHOH	HOHOOH
SynPop (%)	21.4	15.7	14.2	7.2	4.5	4.2	3.4	2.5	1.6	1.5
MPD (%)	19.6	4.3	17.7	2.9	10.0	< 1	6.9	4.7	1.4	1.2

Regarding the time of activities, it can be observed that people in the SynPop leave home earlier on average (09:00 am) than in the MPD (09:45). At the same time, people in the MPD on average return home later (18:45 pm) than the SynPop (17:10 pm). Consequently, a minor discrepancy in the daily rhythm is evident between the two data sets. This offset is underlined by the observation of shorter activity durations in SynPop, with a mean out-of-home activity duration of approximately 6 hours, as compared to 7 hours in the MPD. One possible explanation is that the activity plans generated by actiTopp are based on data from 2017. Possible shifts due to increased home office, for example, are therefore not taken into account in SynPop.

Overall, the data sets show some differences, especially with regard to the timing and complexity of the activities. Despite the different data sources, the various processing steps provide a suitable basis for the following data matching.

2.2. Matching-Process

The goal of this study is to integrate SynPop data with MPD to refine the spatiotemporal localization of trip chains within the SynPop. To do so, the package *MatchIt* in R Studio was employed. An overview of the workflow can be found in figure 1 and 2.

Pre-Processing and Indexing. For the pre-processing, we first removed duplicates from both data sets. Additionally, we standardized the temporal units, ensured consistency in the number of digits in zone IDs, and handled missing data by removing or imputing the values. Further, we generated unique IDs for each generated activity chain to improve computational efficiency. As our objective is to represent an average weekday, we processed the entire temporal span of the MPD to maximize the likelihood of generating a match with the one-day-based Synpop data set. As already indicated, we applied extensive pre-processing on our data which resulted in a focus on mainly the essential information required for the matching. Consequently, it was not strictly necessary to apply indexing for efficiency reasons to reduce the Cartesian product of $n \times m$, between the databases SynPop with size n and MPD with size m.

Data matching. Overall, six key attributes were used to conduct the matching (figure 1 and 2). The first part of the matching process involved the exact matching of attributes related to residential and workplace locations and activity chain patterns. The first two are based on the municipal level. To ensure high-confidence matches, acknowledging the absence of control variables to validate real matches, exact matching was applied. Then, the activity chain IDs were matched also exploiting exact matching. No weighting of these two attributes was applied due to their strictly rule-based nature.

In the second part of the process, the rule-based approach was extended by incorporating statistical computation methods to refine the list of potential matches generated earlier. The Nearest Neighbor matching method with Mahalanobis distance (MAH) was applied to evaluate congruence in the attributes activity duration for the activities ”Work” and ”Other” and the point in time of the first and last activity. The first activity is defined as the starting time of the initial activity after the end of the first home activity, whereas the last activity refers to the time following the final activity, including the last trip back home. The Mahalanobis distance is a method used to assess whether two groups are comparable in terms of specific characteristics. It takes into account the correlations between the variables, as well as the distance of individual data points from the mean (Olmos and Govindasamy, 2015). Literature states that Mahalanobis distance yields more reliable results when applied to data sets with a small number of continuous covariates given its ability to effectively account for the covariance structure. Also, it provides accurate distance mea-

sures without being influenced by multicollinearity (Stuart, 2010). The Nearest Neighbor approach is a method used to identify the closest match for each data point from the other data set, based on the smallest distance (Beyer et al., 1999). Thus, in our case a potential matching partner for a given data point is ultimately the one with the smallest Mahalanobis distance. A 1:1 matching approach was used as the attributes homezone and workzone are based on spatial units. This ensures that the matched zones are geographically close, maintaining a high level of similarity. Furthermore, this method helps to reduce computational time compared to a 1:k matching. Lastly, we also performed the matching with and without the caliper. The caliper ensures that only matches within a specified maximum allowable distance are considered, thereby reducing potential biases. However, this approach also entails a trade-off in terms of information loss, as it imposes an artificial boundary on the accepted distances between two data points. As a result, some potentially suitable matches may be excluded if they fall outside this predefined distance range.

Propensity Score Matching (PSM) serves as an alternative to the Mahalanobis distance. Instead of relying on correlations and means, PSM calculates a propensity score, which quantifies the probability that a given data point belongs to a specific data set, based on its covariates. This score consolidates the examined covariates into a single metric, thus reducing the dimensionality that may impede the matching (Olmos and Govindasamy, 2015). Due to its reduction of covariate information to a single-dimensional score, PSM, despite its widespread use in the literature, has been subject to criticism for oversimplifying the multidimensional nature of the matching process (Kurz et al., 2024). The propensity score is typically computed using logistic regression or a similar statistical technique. In line with Nearest Neighbor matching, matches are then identified by selecting the data point from the other data set whose propensity score is closest to that of the given data point, thereby ensuring similarity in their covariate profiles (Olmos and Govindasamy, 2015).

As seen in figure 1, data pairs were subsequently transferred to the final list once they met all matching criteria. Once added to the final list, these SynPop data points were removed from the further matching process. To allow greater flexibility, the ”with replacement” function was applied to the MPD. In this way, data points from the MPD are used multiple times for matching with the SynPop. This iterative approach enabled maximizing the number of plausible matches identified during the process.

Due to the iterative nature of the matching process, an optimization loop must be defined for the algorithm. Given its computational intensity and the need for efficiency, a greedy algorithm was applied, even though it does not guarantee a globally optimal solution (Vince, 2002). This algorithm operates by iteratively selecting the best match for each data unit step by step. The algorithm continues to run until no more matching data points are available. In this way, the final list contains data points that adhere to the matching rules and can therefore be uniquely assigned. Data points that could not be matched were reviewed and either added to the final matching list or discarded. Reasons for a lack of matching can, among other factors, be attributed to data quality issues.

Assessing balance. Balance analysis plays a vital role in assessing matching quality, as it compares the distributions of the covariates across the matched data sets. This ensures that the matched samples are sufficiently similar, minimizing biases introduced during the matching process. To evaluate the balance of the matching process, standardized mean differences (SMD) and variance ratio between both data sets can be employed. They evaluate whether the distributions of matching attributes are comparable across the matched data sets.

Assessing matching quality. To evaluate the validity of our data matching approach, we lacked a ground truth data set and thus adopted an alternative verification method. We calculated the travel distances between origin and destination zones in the MPD using car matrices and assigned this information to the matched data pairs. We assume for this analysis that the travel distance by car is representative for all modes as no information on the travel mode is available yet. Consequently, each matched person was assigned a daily travel distance. The average daily travel distance was then calculated and compared to the *Mobility in Germany* survey results (Nobis and Kuhnimhof, 2018). A correspondence between the distances indicates that the spatial enrichment of Synpop using MPD is realistic. In a further step, we closely examined the unmatched data points from SynPop to determine whether one or more covariates were responsible for the unmatched portion.

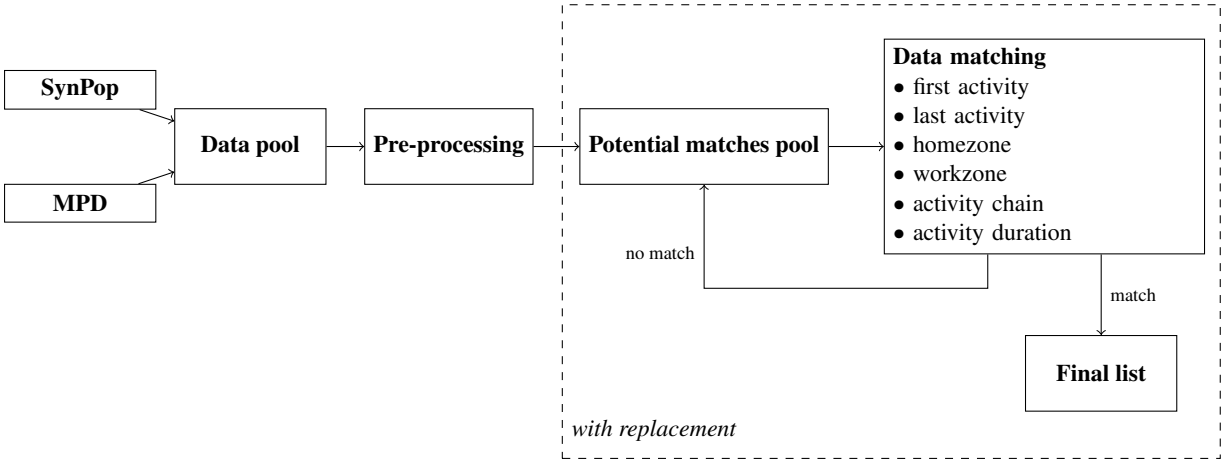

Fig. 1. Schematic workflow of the methodology

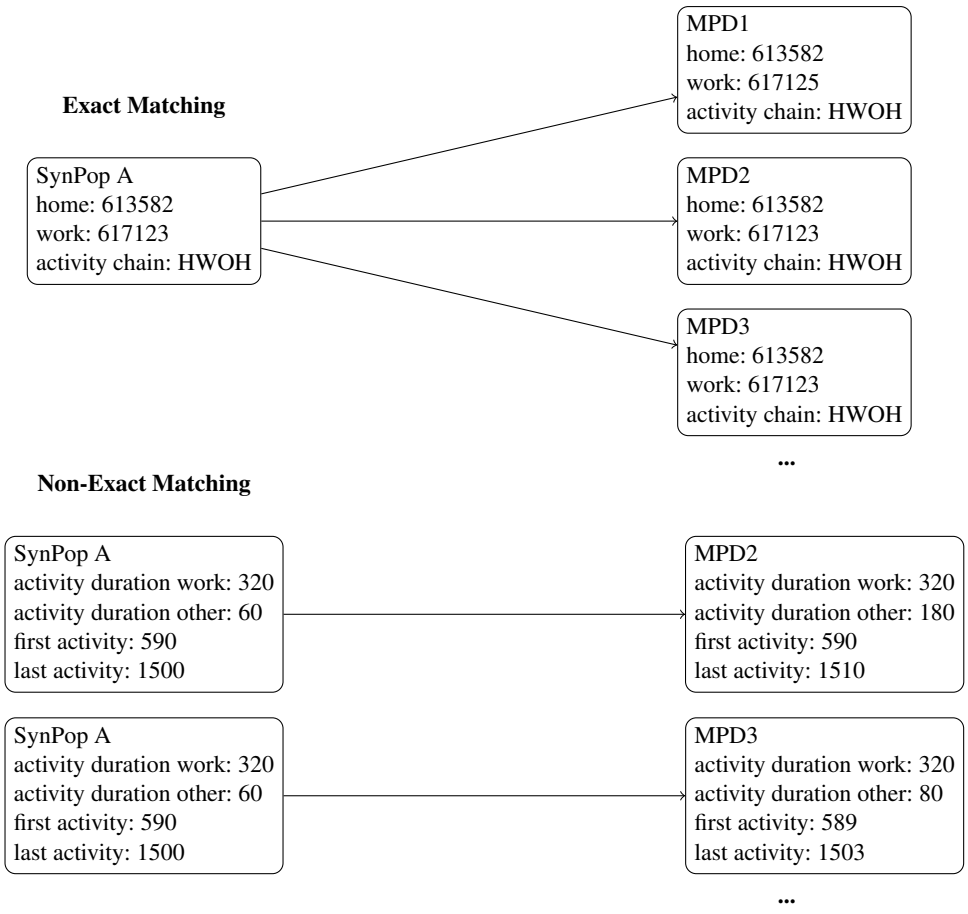

Fig. 2. Schematic workflow based on an example

3. Results and discussion

To maintain a comparison of the balance before and after matching, SMD and variance are commonly used. In this way, improvements in the values for each covariate can be used to assess the quality of the matching process: after matching, unmatched data points, such as those due to extreme values, may be excluded from the data set, which can lead to improvements in both SMD and variance. Consequently, both of these metrics are employed before and after matching for a balance assessment. It is crucial to mention that this applies to non-exact covariates as exact matches already guarantee balance for these variables, meaning there is no need for further assessment of balance.
Before matching, differences in means, SMD and variance ratio (table 2) can be found between SynPop and MPD. All SMD for last activity, first activity and activity duration reveal a moderate inequality before matching. Whereas only 62% of the variance of the first activity in MPD can explain the variance in the covariate first activity in SynPop, 75% can be explained for the covariate activity duration. This shows the greater data variation for the first activity and activity duration in MPD while variation is higher for the last activity in SynPop. In the MAH, a total of 558,308 (68.13%) out of 819,442 data points from SynPop and 195,343 out of 1,454,824 (13.43%) data points from MPD were successfully matched (table 3). No data points were discarded including data that was not used for the matching process. The mismatch of unmatched data points in the MPD is not of particular concern, as it merely serves as a pool of potential data matches, with the primary focus placed on the completeness of matches and the improvement of the data quality of Synpop. Subsequently, the balance between the data sets was evaluated to assess the quality of the matching.

Table 2. Output for All Data before matching

Balance Metrics	Means SynPop [min]	Means MPD [min]	SMD	Variance Ratio
First activity	539.80	590.97	-0.31	0.62
Last activity	1,029.77	1,107.89	-0.37	1.19
Activity duration	361.52	408.57	-0.21	0.75

Table 3. MAH and PSM summary for matched sample sizes (with caliper in brackets)

	SynPop	MPD
All	819,442	1,454,824
Matched MAH	558,308 (411,222)	195,343 (160,731)
Unmatched MAH	261,134 (408,220)	1,259,481 (1,294,093)
Discarded MAH	0	0
Matched PSM	558.308 (495,774)	258.912 (244,072)
Unmatched PSM	261.134 (323,668)	1.195.912 (1,210,752)
Discarded PSM	0	0

Post-matching all three variance ratios (table 4) have improved, in other words are closer to equal variances. Whereas the range of SMD before matching is from -0.37 to -0.21, it has improved to -0.12 to 0.01 (MAH). An improvement in the variance ratio can also be observed as variance ratio for first activity has improved from 0.62 to 1.07, for last activity from 1.19 to 1.03 and activity duration from 0.75 to 1.07. Thus, the matching process has contributed to making the distributions more comparable. An interesting aspect is the comparison of MAH with and without the use of a caliper $c = 0.5$ (table 4). While fewer matches were identified when using a caliper, the comparison of pre- and post-matching balance metrics demonstrates that the matching process with a caliper resulted in some statistical improvement compared to the method without a caliper. For instance, variance ratio for the first activity in MAH with caliper refined minimally compared to MAH.

The PSM demonstrated that a total of 558,308 (68.13%) out of 819,442 data points from Synpop and 258,912 out of 1,454,824 (17.8%) data points from MPD were successfully matched. The balance metrics after matching indicated

Table 4. Output for Matched Data (with caliper in brackets)

Balance Metrics	Means SynPop	Means MPD	SMD	Variance Ratio
First activity MAH	569.01 (588.57)	577.62 (590.91)	-0.05 (-0.01)	1.07 (1.03)
Last activity MAH	993.45 (976.56)	1,018.14(984.08)	-0.12 (-0.04)	1.03 (1.01)
Activity duration MAH	304.51 (276.08)	302.97 (267.69)	0.01 (0.04)	1.07 (1.05)
First activity PSM	569.01 (578.19)	582.29 (590.68)	-0.08 (-0.08)	1.05 (1.05)
Last activity PSM	993.45 (994.05)	1044.40 (1035.70)	-0.24 (-0.20)	0.92 (0.88)
Activity duration PSM	304.51 (297.10)	278.69 (257.28)	0.18 (0.18)	0.88 (0.93)

that the SMD were closer to 0, suggesting a stronger alignment of the covariate distributions between the groups after matching. Similarly, the variance ratio indicated that the variability has improved in all three covariates before and after matching. Again, various balance metrics improved slightly when applying the caliper compared to without the caliper (table 4). However, these improvements remain in the two- or three-decimal range and can therefore be considered minimal.

As previously mentioned, the introduction of a caliper resulted in some improvements in the balance metrics. This may be partly due to the fact that the caliper, through its tolerance threshold, only matches very similar data points. As a result, pairs with larger distances (exceeding the threshold) are excluded, which can improve the balance metrics of the overall covariate set, however, potentially discarding valuable matches and information. This also implies that potential outliers in a covariate group are less likely to be matched, thereby increasing the homogeneity of the covariate set. However, since the aim of this paper is not to explain causal relationships between the two data sets but rather to focus on the completeness of the matching process and to prioritize the preservation of sample size, the method without a caliper appears to be more suitable. This trade-off between balance and sample size reflects the priorities of this paper, which emphasizes the completeness of the matching process while maintaining an acceptable level of balance. In line with existing literature, the MAH matching process is particularly well-suited for data sets with a limited number of continuous covariates, as it effectively minimizes the differences between both groups of each covariate by directly accounting for the covariance structure. This aligns closely with the structure and goals of our data set. In contrast, the PSM method, while widely used, has shown a tendency to oversimplify the multidimensional relationships between covariates by reducing them to a single scalar propensity score. This may result in a less nuanced matching process. Thus, considering the specific characteristics of the data set, the balance metrics achieved, and the methodological strengths of MAH, it can be concluded that MAH offers a more methodologically sound approach to matching in this context. One of the main intentions of the matching process is to create realistic spatio-temporal behavioral patterns for a synthetic population. As mentioned in the validity section, to ensure appropriate spatial characteristics the total distance traveled is assessed. The average total distance traveled per person and day in the synthetic population is 25.7 kilometers excluding the non-mobile individuals. This is less than the respective survey result of the *Mobility in Germany* (Nobis and Kuhnimhof, 2018) estimating 42 and 47 kilometers per day for urban and rural spatial typologies that are comparable to the area of application. The lower value is due to the MPD covering only trips within the considered area and thereby excluding long-distance trips. We expect closer estimates when trips within the same area are compared.

Lastly, it is crucial to analyze non-matches in the SynPop for validity reasons. Table 5 shows the shares of individuals assigned with workzone (employed individuals) and individuals without workzones (including children, retirees, unemployed individuals). In the unmatched data, only 9.4% of individuals lack a workplace compared to 58% in the overall data and 80.3% for the matched data. This indicates that exact matching of the relation between the place of work and place of residence for employed individuals leads to a high number of non-matches if the exact combinations do not appear in the MPD. However, the home and workzones of the non-matched data are distributed evenly across the area. Based on this information, we ran the MAH again, excluding the workzone as a corvariate that requires exact matching. The results reveal the improvement not only in overall matches (91,1% of the SynPop) but also in the closer alignment of the workzone and no workzone distribution in SynPop, as shown in Table 1. This highlights the need for a more precise harmonization of the suitability of covariates with respect to the available data beforehand. Additionally, work activities lead to more complex activity chains, since a third activity is added alongside ”Home”

and ”Other”. HWH and HWHOH are the most common activity chains in the unmatched data of SynPop (20% and 15%). In the matched data set, these chains only occur in 11% and 3% of the cases. Generally, longer activity chains can be observed in the unmatched data set, which leads to a more complex matching of the data points. This could also be explained by the fact that short trips are underrepresented in the MPD, as they do not exceed the detection threshold and therefore more complex SynPop activity chains do not find a match in the MPD. The ten most frequent activity chains in the matched SynPop show an average of 4.4 activities, while the figure for the data set of non-matched SynPop is 5.1 activities. On the one hand, this could be attributed to the fact that SynPop generation creates complex activity chains that do not occur in the MPD with the same frequency. On the other hand, it is also conceivable that the MPD, due to its 25% coverage of the German market or the data processing methods applied, does not provide a complete or detailed representation of all possible activity patterns. Certain travel behaviors that are more commonly represented in the synthetically generated SynPop data may be absent from the MPD, leading to a discrepancy between the two data sets. The improvement in balance metrics for the first and last activities, as well as activity duration, indicates successful matching for these covariates. Additionally, the descriptive statistics for the matched and non-matched Synpop data sets show minimal differences, and the distributions for homezone remain consistent before and after matching. This suggests that the primary reasons for non-matching are likely the workplace and activity chains.

Table 5. Share of individuals with and without workzone

	No Workzone	Workzone
SynPop	58%	42%
SynPop - Matched	80.30%	19.70%
SynPop - Unmatched	9.40%	90.60%
SynPop - Matched (MAH Matching without workzone)	60%	40%
SynPop - Unmatched (MAH Matching without workzone)	34%	66%

4. Conclusion

Our goal of enriching the synthetically generated population with spatial from mobile phone data was partially successfully achieved through the presented data-matching process. Although the matching was successful with regard to the non-exact covariates, as there were improvements in balance, discrepancies were found in the distribution of the exact covariates. These differences must be critically scrutinized, e.g. due to different distributions in the underlying data sets or possible limitations in the matching algorithm.

Building on existing literature, we implemented a matching process that includes several steps of data preparation, along with the core matching and validation procedures. Initially, the synthetic population was generated using conventional methods. In comparing these data sets, we established six criteria that served as matching covariates. To generate as many plausible matches as possible, a hybrid approach combining rule-based and probabilistic matching methods was applied. The alignment of home and work locations as well as activity chain IDs for each data point from both data sets required exact matching. In contrast, for first and last activity per day and activity duration, summed over a day, matching was based on the two different distance metrics Mahalanobis and Propensity Score, which reflects the similarity between units in terms of these covariates. Subsequently, both distance metrics were run again with a caliper. A greedy algorithm was employed, despite its trade-offs compared to an optimal algorithm as it shows a higher computer efficiency. Eventually, the algorithm stopped running iteratively until no more data points could be matched. Based on the analysis of balance and validity, we found that the Nearest Neighbor matching with Mahalanobis distance provided a more reliable matching for our data. In total, 68.13% data points were matched from the synthetic population data pool. The subsequent analysis of balance revealed that the MAH method appears to be slightly more suitable for our data. As we were lacking a ground truth data set, we used distance traveled per day as an indicator. This revealed a difference between the travel distances in our data and those reported in the *Mobility in Germany* survey, highlighting the need to revisit the comparison by focusing on trips limited to the same area to achieve more accurate alignment. Approximately 31.87% of the data could not be matched in the first step.

After comparing the unmatched and matched data points from SynPop in an analysis, we observed a discrepancy between the original SynPop data set and the matched SynPop data, particularly in terms of the workzone and activity chain complexity. Therefore, it is essential that covariates are carefully selected and tailored to the data sets being matched. Consequently, it may be advisable to revisit the unmatched data points in a subsequent matching step by relaxing the matching criteria. Specifically, certain covariates could be removed to explore whether these adjustments facilitate additional matches. It is crucial, however, to ensure that only covariates that are not critical in subsequent agent-based modeling steps are removed. For instance, removing the ”workzone” covariate could be a viable option, as this variable can be synthetically generated during later modeling stages. By following this approach, the 68% of data points successfully matched in the initial process could potentially be augmented with additional matches from the second step, thereby enhancing the data set for use in model development.
For future work, we recommend further comparison of different distance measures to achieve the most efficient and plausible data matching. Also, using two data sets from the same time period would improve the comparability of the results. As previously mentioned, distance measures have varying advantages and disadvantages, which may have different levels of significance depending on the data set. To minimize this trade-off, it is advisable to adopt a more statistically grounded approach and conduct a comparison based on the balance and validity of individual distance measures. In particular, the ratio of rule-based to statistical approaches could be of interest. Furthermore, the influence of the weighting of individual covariates as well as the definition of a data-derived caliper threshold is an important aspect for future research. Finally, comparing greedy and optimal algorithms could be valuable to assess the stability of the data matches.

Acknowledgements

This research was funded by the German Federal Ministry of Transport (grant number 45AVF2B021).

References

Beyer, K., Goldstein, J., Ramakrishnan, R., Shaft, U., 1999. When is “nearest neighbor” meaningful?, in: Database Theory—ICDT’99: 7th International Conference Jerusalem, Israel, January 10–12, 1999 Proceedings 7, Springer. pp. 217–235.
Calabrese, F., Diao, M., Di Lorenzo, G., Ferreira Jr, J., Ratti, C., 2013. Understanding individual mobility patterns from urban sensing data: A mobile phone trace example. Transportation research part C: emerging technologies 26, 301–313.
Christen, P., 2012. The data matching process. Springer.
Demissie, M.G., Phithakkitnukoon, S., Kattan, L., Farhan, A., 2019. Understanding human mobility patterns in a developing country using mobile phone data. Data Science Journal 18:1, 1–13.
Hakak, S.I., Kamsin, A., Shivakumara, P., Gilkar, G.A., Khan, W.Z., Imran, M., 2019. Exact string matching algorithms: survey, issues, and future research directions. IEEE access 7, 69614–69637.
Hilgert, T., Heilig, M., Kagerbauer, M., Vortisch, P., 2017. Modeling week activity schedules for travel demand models. Transportation Research Record 2666, 69–77. URL: https://doi.org/10.3141/2666-08, doi:10.3141/2666-08.
Kurz, C.F., Krzywinski, M., Altman, N., 2024. Propensity score matching. Nat Methods 21, 1770–1772.
Li, Z., Ning, H., Jing, F., Lessani, M.N., 2024. Understanding the bias of mobile location data across spatial scales and over time: a comprehensive analysis of safegraph data in the united states. Plos one 19, e0294430.
Nagels, J., Wu, S., Gorokhova, V., 2019. Deterministic vs. probabilistic: best practices for patient matching based on a comparison of two implementations. Journal of Digital Imaging 32, 919–924.
Nobis, C., Kuhnimhof, T., 2018. Mobilität in deutschland- mid: Ergebnisbericht .
Olmos, A., Govindasamy, P., 2015. Propensity scores: a practical introduction using r. Journal of MultiDisciplinary Evaluation 11, 68–88.
Smoreda, Z., Olteanu-Raimond, A.M., Couronné, T., 2013. Spatiotemporal data from mobile phones for personal mobility assessment, in: Transport survey methods: best practice for decision making. Emerald Group Publishing Limited, pp. 745–768.
Stopher, P., FitzGerald, C., Xu, M., 2007. Assessing the accuracy of the sydney household travel survey with gps. Transportation 34, 723–741.
Stopher, P.R., Greaves, S.P., 2007. Household travel surveys: Where are we going? Transportation Research Part A: Policy and Practice 41, 367–381.
Stuart, E.A., 2010. Matching methods for causal inference: A review and a look forward. Statistical science: a review journal of the Institute of Mathematical Statistics 25, 1.
Vince, A., 2002. A framework for the greedy algorithm. Discrete Applied Mathematics 121, 247–260.

Wang, F., Chen, C., 2018. On data processing required to derive mobility patterns from passively-generated mobile phone data. Transportation Research Part C: Emerging Technologies 87, 58–74.
Wilson, J., 2004. Measuring personal travel and goods movement. Tr News 234, 28.