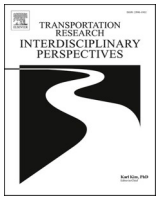
Transportation Research Interdisciplinary Perspectives 24 (2024) 101078

Contents lists available at ScienceDirect

Transportation Research Interdisciplinary Perspectives

journal homepage: www.sciencedirect.com/journal/transportation-research-interdisciplinary-perspectives



A case study of checking national household travel survey data with machine learning

Lisa Ecke [a,*], Miriam Magdolen [a], Sina Jaquart [a], Robin Andre [a], Peter Vortisch [a]

[a] Institute for Transport Studies, Karlsruhe Institute of Technology (KIT), Kaiserstrasse 12, 76131 Karlsruhe, Germany

ARTICLE INFO

Keywords:
Big data
Data validation
Data quality
German Mobility Panel
Machine learning

ABSTRACT

In recent years, machine learning techniques have been increasingly tested and applied to physically collected data to optimize the processes. In this paper, machine learning is used to check travel survey data of the German Mobility Panel (MOP). In the MOP, verified and raw data have been available for several decades, on which algorithms can learn the practices of human checking routines. By using machine learning, the algorithm is expected to learn the checking patterns from the past and thus support the data checking of new datasets. To this aim, several algorithms are applied and tested. The presented model framework supports the identification of blatant deficits in the reports at the individual and trip levels. The neural network (NN) shows the most promising results as it decreases the number of data samples checked. The checking effort can be reduced by 20.4 % at the individual trip level. This work shows that machine learning can support the data checking process in the MOP at various levels, thus leading to significant time reduction.

1. Introduction

The provision of truthful travel survey data is a high responsibility because decisions for future investments are made on it. Thus, the data quality and completeness require special attention (Hubrich et al., 2018). When assessing data quality, a distinction must be made between sampling and non-sampling errors. Sampling errors occur when the sample itself is biased. However, such errors can be corrected by weighting the data and are not the focus of the following study. Non-sampling errors include all problems that are not directly related to the sampling but endanger the data quality (Aschauer et al., 2018; Bonnel et al., 2015; Hubrich et al., 2018). Ensuring data quality is, however, however always time-consuming and, therefore, costly.

In the German Mobility Panel (MOP), a Germany-wide national household travel survey (NHTS), around 3,800 people are surveyed annually about their everyday travel (Ecke et al., 2021). For this purpose, participants keep a so-called trip diary for one week to report all their trips. In the 1990s, a methodology for data validation was developed and has been consistently applied to manually check for non-sampling errors and identify individuals with glaring reporting deficiencies. The methodology has been continuously developed further. Besides various algorithm-based checks, the data assessment of non-sampling errors has also been done by trained staff. Through the continuous application of the checking routines in all years of data collection, the data was made available to planners with consistent data quality. Comprehensive data documentation of the checking process is available, which allows in-depth insights into the checking process of the data. However, with approximately 70,000 trips per survey wave, this process is very time-consuming and thus costly, as the trained staff must check every person and trip.

In light of the above, machine learning techniques seem promising support for the data checking and dropout identification process for the MOP data. The previously checked data can be seen as large labelled datasets, which can serve as a basis for supervised learning techniques. The machine learning algorithm is expected to learn the checking patterns from the past to support data preparation and checking in newly collected datasets. Therefore, this paper will address the questions to what extent and how successfully machine learning can support the data preparation and checking process of the MOP.

The paper is structured as follows: First, a literature review draws a picture of the data checking processes of travel survey data and to what extent machine learning is already used for data validation. Second, the data used as well as the past data checking process are described. Third, machine learning techniques are applied, tested and evaluated at

* Corresponding author.
E-mail addresses: lisa.ecke@kit.edu (L. Ecke), miriam.magdolen@kit.edu (M. Magdolen), sina.jaquart@gmx.de (S. Jaquart), robin.andre@kit.edu (R. Andre), peter.vortisch@kit.edu (P. Vortisch).

https://doi.org/10.1016/j.trip.2024.101078
Received 13 April 2023; Received in revised form 6 October 2023; Accepted 21 March 2024
Available online 23 March 2024

different steps during the data checking processes of the MOP. Next, the limitations of the presented methods are described. Last, the impact of using machine learning is investigated and discussed. This study ends with a conclusion and an outlook for further research.

2. Background

In this section, we provide an overview of data validation of travel survey data in the literature, introduce the MOP survey design and describe the data assessment methods of the MOP data in the last decades.

2.1. Literature

Machine learning is receiving increasing attention in the transportation sector. For example, methods such as the neural network (NN) are used to predict usage frequencies and vehicle preferences (Hu et al., 2022). Machine learning is, for example, also used for the prediction of shared-car use. A study by Wang et al. found that carsharing operators accurately predict the station-level shared-car use and optimally identify the best locations for stations, thus maintaining the operational efficiency of carsharing programs (Wang et al., 2021). In particular, mobile device location data provide a comprehensive database to answer questions about individual travel. By using machine learning, this data can now be efficiently analyzed to better predict and model population inflow, for example (Hu and Xiong, 2023).

With the increasing number of machine learning applications whenever data is analyzed, the question arises if machine learning can also help to check survey data. When it comes to studies that apply machine learning in the context of travel survey data, two main applications can be identified: transport mode recognition (e.g., Feng and Timmermans, 2013; Wang et al., 2018; Zhou et al., 2016) and trip purpose identification (e.g., Meng et al., 2017; Montini et al., 2014). All these studies have in common that they are based on automatically collected data, e.g. GPS trajectories, and not revealed preference data or other types of survey data. Furthermore, the literature highlights the general challenge of using such data for prediction due to complex nonlinear temporal dynamics in the travel of individuals as they show high-dimensional and non-negligible impacts from various external factors (Hu and Xiong, 2023).

Von Behren et al. (2020) apply machine learning to panel survey data on everyday travel to segment people based on visualization of weekly travel. The application does not focus on data checking but on the segmentation of homogenous groups, which is also essential for understanding travel behavior. However, these studies generally focus on generating and understanding travel data and less on applying machine learning techniques to check and assess data quality.

Depending on the data collection methodology, the effort to check the data and ensure high data quality varies. For example, during online surveys, the provided data can be checked in real-time and warnings can be issued to the participants if data is likely to be wrong or missing (Aschauer et al., 2021). The NHTS in the U.S. uses this method to support participants while reporting and thus to achieve higher data quality (NHTS, 2018). Surveys using paper and pencil as data collection methods do not allow for real-time assessment and are costly as errors must be identified after data collection (Couper, 2011).

When working with NHTS data, it is challenging to evaluate participants who do not report any trips, either out of lack of motivation (dropout) or because no trips were made (immobile). It becomes even more complicated if only some trips are not reported because these 'abnormal' could probably reflect authentic travel patterns and cannot be directly adapted in the dataset. The phenomenon of item-nonresponse is examined by Aschauer et al. (2021). By contacting the participants promptly after the first data check, follow-up data completion and validation were possible. It was found that trips were mainly underreported on days when many trips were made. Similar results were found by Wittwer and Hubrich (2015). The decreasing motivation of participants is even more critical in longitudinal surveys, in which participants are asked to report several times and/or over a more extended period. Some participants start motivated at the beginning of the survey and lose motivation, and the data gets biased or erroneous over the reporting period (Chlond et al., 2013; Wirtz et al., 2013). Kitamura and Bovy (1987) found that participants who reported accurately in the first wave of a survey were also likely to do so in the second wave. However, identifying unplanned dropouts is challenging and has been investigated by De Haas et al. (2022) and Kuhnimhof et al. (2006). Madre et al. (2007) and Armoogum (2014) suggest that many participants are unplanned dropouts in travel surveys.

In summary, numerous studies use ML for predictions in the transportation sector. However, the areas of application are limited to technically collected data. According to the authors' research, data checking of travel behavior data captured with surveys does not exist so far. The reasons are mainly the lack of ground truth data and the high variability of individual travel behavior. A research gap is identified regarding the lack of experience with ML in revealed preference data checking. In the following, the case study investigates how to automate the checking process of travel survey data, exemplarily tested on the MOP.

2.2. German Mobility Panel

The German Mobility Panel (MOP) is an annual NHTS. Since 1994, it has been carried out on behalf of and funded by the German Federal Ministry for Digital and Transport. The market research firm KANTAR is responsible for the fieldwork (i.e., recruitment and data collection) and the Institute for Transport Studies of the Karlsruhe Institute of Technology (KIT) is in charge of the survey's design and scientific supervision (Ecke et al., 2021; Jödden and Führer, 2021). The data collection takes place in the fall. Participants are asked to fill in a trip diary over seven consecutive days. The trip diary provides information about all trips during this week (distances, means of transport, trip purposes and departure and arrival times). Participants also indicated irregularities, for example, whether they were ill or on vacation. Furthermore, sociodemographic information about the participants and households is asked for.

The annual sample size is 1,500–1,800 households with 2,600–3,100 persons aged ten years and older. The MOP is designed as a rotating panel. Participants are asked to report their travel behavior in three consecutive years. A new cohort of first-year reporters replaces a portion of the sample that retires every year after three years of reporting. Every year, the data is stored in three datasets: household, person, and trip dataset. All datasets can be linked via key variables. More information can be found on the project website (Ecke et al., 2022).

This paper uses five years of data (2015–2019) with three datasets per year (household, persons, trips). The sample size of each dataset is displayed in Table 1 (individuals can report in up to three years).

2.3. Data checking process of the last decades

Data checks are carried out within the data assessment to ensure high and consistent data quality. The three-step assessment aims to identify people with blatant deficits in their reports (dropouts), and to detect non-sampling errors (e.g. missing values). The procedures have been

Table 1
Sample size of the MOP surveys 2015–2019.

Survey year	Persons	Households	Trips
2015	3,774	1,843	64,418
2016	3,643	1,776	67,065
2017	3,867	1,881	71,977
2018	3,835	1,868	73,041
2019	3,872	1,864	72,216

used since the start of the survey in 1994.

Dropout identification occurs for the first time during the pre-checks so that the relevant participants are removed from the datasets. The data checks are divided into preliminary checks and individual checks. The preliminary checks examine the personal data based on predefined rules. In addition, another search for dropouts occurs during the separate checking of the trips. In this step, persons are also removed if the data are incomplete and unusable for further analyses. In the individual checks, the daily travel and activities of the persons are checked over one week. Software is used for this purpose, described in the next section. Due to the annual implementation and the almost unchanged survey design, the data structure has been nearly unchanged over the last decades. The documentation of the data checking process (checking rule set) and the checked data itself over the past decades make it possible to provide a database on which an artificial intelligence-based algorithm can be trained to learn the data assessment from historical data and apply it to new data.

2.3.1. Graphical diagnosis of individual travel behavior (GraDiV)

Since the start of the MOP, the so-called GraDiV software is used for the trip data checking. GraDiV is used for the individual checks of each participant (e.g., check for completeness, identification of not completed trip chains etc.), which is described in detail in Ecke et al. (2024). Results based on the data checked with GraDiV are presented in (Ecke et al., 2021; Jödden and Führer, 2021). In the tool, the activity and trip schedule of the week are visualized and thus allow to see implausibility in the data reported directly. The checking processes of the past years are documented. The data obtained (raw data and the data after data checking with GraDiV) is suitable as training data that allows algorithms to learn the nuanced human judgment of the data checking process.

3. Methodology

In this paper, applications of machine learning methods for data checking are evaluated and their functionality is tested. Based on the manual data assessment of the past (GraDiV), it is investigated to what extent the nuanced human judgment can be learned and applied by algorithms at different data levels (here: individual and trip level). The basic idea is that algorithms learn from a training dataset which is based on manually checked data of the past and is then applied to a new raw dataset. In this process, it is first tested whether unplanned dropouts can be identified in the data by machine learning methods, i.e. to what extent these “bad risks” in the data can be detected. For this, we apply several different machine learning methods (neural network, decision tree, random forest and support vector machine). Furthermore, it is also tested whether incorrect trip data can be identified by machine learning methods and whether these can be corrected automatically.

3.1. General concept

For this work, a three-stage framework was developed based on the stages of the (manual) checking process from the past. This enables comparing both processes (manual vs. machine learning) at all levels. Fig. 1 shows the stages for which machine-learning approaches are implemented and further tested. The first step is to look for dropouts during the pre-checks (1). For this purpose, only participants who have reported no or only a few trips are relevant. Identified dropouts (including their reported trips, if existent) are removed from the person dataset. Then, as part of the individual checks (2), all remaining participants are considered.

As a third step, the individual trips of the trip dataset are checked for implausibilities during the individual checks (3). Trips that may need to be manually checked afterward are marked as such. The model implementations, data preparation, and model performance are presented in section 3.3.

All models designed in this work are implemented using the Python 3.8.0 release. For data preparation, the libraries Pandas and NumPy are used. Machine learning and deep learning is performed using TensorFlow, Keras, Scikit-Learn and TensorBoard.

3.2. Model framework for dropout identification in pre- and individual checks

This section highlights the applicability of machine learning to support the identification of dropouts. Identifying dropouts as early as possible in the preliminary checks is desirable, as this means that individual’s reporting data will only go through a few steps of the checking process. As a result, less time and effort are spent on less meaningful data. Remaining participants are again screened for possible report dropout during the individual checks. It is investigated how the algorithms distinguishes dropouts (participants who drop out of the survey prematurely) from participants with major deficits in reporting quality and from participants with justified deviations from expected behavior (e.g., due to illness). In the pre-checks and the individual checks, the goal is to detect faulty persons (dropouts). The algorithms do not correct or supplement data.

3.2.1. Model implementation and data preparation

The data preparation for dropout identification during the pre-checks ((1) Fig. 1) and individual checks ((2) Fig. 1) is similar. The

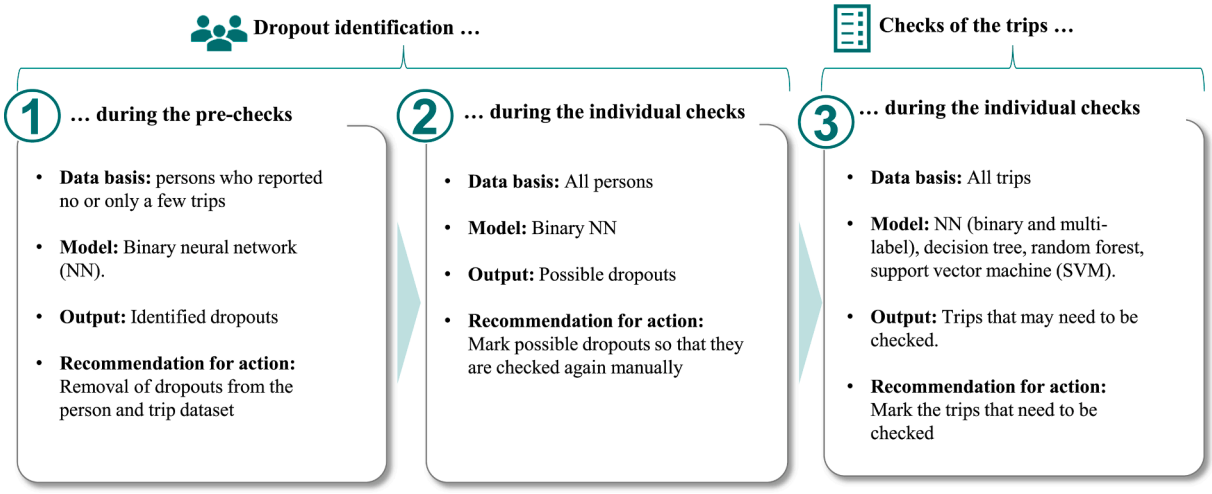


Fig. 1. Connection of the application of the models for dropout identification and individual trip checks.

sociodemographic data of the participants and aggregations regarding their reporting behavior serve as the data basis. In the individual checks, however, the individual trips made by each person are also taken into account. Neural Networks (NN) are implemented for both pre-checks and individual checks for the identification of dropouts. We use NNs because there is comparatively little systematic knowledge in the datasets and a large amount of imprecise information that needs to be processed, and NNs can handle this better than other algorithms (Dreyfus, 2005).

First, it must be ensured that the input layer dimension has the appropriate number of input features. The input layer of the NN for dropout identification during pretesting receives nine features (age, presence of a mobility impairment, occupation, household type, number of days reported, number of trips reported, number of days with sickness, number of vacation days, and number of days reporting abnormalities such as a business trip, see Appendix, Table 5). The features are taken directly from the person and trip datasets. Including the dummy variables created by One Hot Encoding, these nine features are represented by 20 input columns. For the individual checks, the dimension of the input layer is 25 because reported trips (five columns) are considered in addition to sociodemographic data. A fully connected network structure with four layers is used in both NNs, i.e.; there are two hidden layers. The two NNs (for the pre-checks and for the individual checks) are each trained using both single datasets from 2015 to 2019 and aggregations of datasets from several years.

Regarding the question which data should be used to train the final model, there is a particular trade-off to consider: Datasets from as many years as possible should be used so that the model can learn with many data points. However, the NN will become confused if many data points are added that do not reveal a clear pattern. In addition, the consistency and comparability of the results increase if the dropout identification during the pre-checks and during the individual checks are based on the same years.

Only individuals who reported no or few trips within the one-week reporting period and their household members are considered in the pre-checks by the algorithm. However, a small number of reported trips alone is not sufficient as a characteristic for dropout identification. For example, the individual could be an older person who is being cared for by another household member because they may be mobility-impaired themselves. In this case, the person should be considered in later analyses and thus not identified as a dropout.

It is essential to note the distribution of the labels. The confusion matrix is presented in the Appendix (Fig. 3). It can be seen that 72.22 % of the data points are correctly positively classified and 21.3 % are correctly negatively classified. In this case study, the non-dropouts are the “abnormal” class whose detection needs to be optimized. This is because those labeled as dropouts are removed directly from the person dataset. In addition, their trips are deleted from the dataset for the individual checks of the trips as it is not recommendable to keep bad training data for the sake of quantity. However, since each participant’s trip diary has a high value for the survey, unjustified deletion of a participant’s data should be avoided. In this case, it is worse to label a person as a dropout who is not, than not to label a person as a dropout who could be a dropout. If the latter case occurs, there is still the possibility to identify this dropout during the individual checks. Generally, it is a trade-off between data quality and quantity. Before applying the model, the categorical features such as the presence of a mobility constraint, occupation, and household type are each One Hot Encoded. In addition, the dropout class is weighted more heavily during the pre-check. It aims to make the model also represent the underrepresented class well. Thus, the weights for the non-dropouts and the dropout class are 0.62 and 2.57 for the pre-check and 0.51 and 24.25 during the individual check.

For the dropout identification during the individual checks, the reported trips are considered in addition to the sociodemographic data. Seven features are constructed for dropout identification during the individual checks, each indicating the number of reported trips per reporting day (one per day; see Appendix, Table 6). The features are designed based on the assumption that dropouts initially report regularly but that the number of reported trips per day decreases as the reporting week progresses.

Six features are considered: mobility restriction, occupation, household type, number of days reported, number of trips reported, and the number of days on which abnormalities are reported. In combination, the reconsideration of these features leads to new patterns.

However, the decoding of the labels differs from the process used during the pre-checking: In this step of dropout identification, all possible dropouts must be detected. Therefore, in this case the class of dropouts is given the label 1 and their recall, i.e. the percentage of detected dropouts with respect to the number of actual dropouts, has to be maximized since no non-dropout should be mistakenly removed from the dataset. Subsequently, the categorical features are also One Hot Encoded.

The training/test ratio of the model is 80/20, while *binary_crossentropy* is used as a loss function. We use three hidden layers and 100 neurons per hidden layer to control the learning process. The learning rate is 0.001, and the batch size is 100.

3.3. Model framework for individual trip checks

The following framework is designed to check the trip diaries. The goal is to classify the trips into two classes: The class of correctly reported trips (Ok) and the class of trips that might need further checking (P). Trips of class P are forwarded for further manual checks because it was found that it is not possible to fully automate the checking process with the given methods.

3.3.1. Model implementation and data preparation

In the first step, labels for the individual trip data are created by comparing the raw trip data with the data after the manual checks by the staff. The labels are defined by identifying how the trip was changed during the manual checks. The checking rules are listed in the GraDiV manual.

First, a general binary label determines whether a trip’s purpose, distance, mode, or start or end time was changed. This label does not yet contain any information about which variable was adjusted. Therefore, a binary label is added for whether this characteristic was manually changed for each trip characteristic. It is also labeled whether a trip is deleted or added.

Features are selected based on their importance to the performance of the model based on the VIANN method described by Kralj Novak et al. (2019). An alternative choice of determining feature importance would be SHAP (Lundberg and Lee, 2017). However, the results of the VIANN method are further used and presented in the Appendix (Fig. 4). In this method, the variance of the weights is measured to obtain the relative variable importance for NNs. During the training phase, the weights of an NN are changed until the model reaches its final state. While testing the model, it was found that not all the features are necessary for the model’s performance, so we could eliminate them and use only the relevant features. The features and their importance are described in more detail in the Appendix. After several trainings it becomes clear that the features *startzeit, endzeit, x_dauer and x_geschwindigkeit* have only a small importance for the NN despite the continuous values. In addition, the features *x_zweck_w + 1, x_hh_gleiche_angabe und x_vm1_w + 1* are of very little importance. As a result, these features are not used to train the final model, as they unnecessarily add complexity and thus degrade the model’s performance.

The categorical features are One Hot Encoded to ensure that the models consider the relevant features. It is important to optimize the classification of the trips to support the processor in the individual checks as effectively as possible. For this purpose, the performance of five models is compared. The dimension of the input layer is equal to the

number of features after decoding and is thus 92. The batch size is 400, the number of training epochs is 8, and the optimizer Adam is used. With more epochs, an overfitting effect occur, i.e., the model memorizes the training examples instead of recognizing generalized patterns. As a result, the NN cannot classify new, unknown examples well. We did not consider employing techniques to mitigate overfitting, such as incorporating dropout layers, early stopping, or skip connections, because the presented work is a case study. However, this might be future work.

The weights of the classes vary depending on the dataset and on how frequently each class occurs in a given dataset. The aggregated dataset from 2015, 2016, 2018, and 2019 is used to train the final model. The weights for training are 0.54 for the Ok class of trips that do not need to be checked and 6.93 for the P class of trips that may need to be checked. In the aggregate dataset, nearly 13 times as many Ok class trips occur as P class trips. The weights for the two classes apply to the trained models and are initialized randomly.

Furthermore, the performance of the four models NN, Decision Tree, Random Forest and SVM is considered. All models are each trained with the datasets of the individual years and aggregated datasets from several years. The performance measures are based on an average of 20 training repetitions to ensure replicability.

4. Results

4.1. Performance of the NN for the individual checks

Table 2 summarizes performance measures to evaluate the NN for dropout identification during the pre-checks. The results are based on an average of 20 replicates, ensuring the approximate results' approximate replicability. When comparing the performance of the models between the different years, it is clear that the model has a challenging time seeing a pattern in the 2015 dataset. Based on the 2015 data, the model has a recall of only 0.54 for the non-dropout class. In subsequent years, the recall is always at least 0.73; in 2016, 2017, and 2019, it is 0.86 or higher.

The performance of dropout identification during the individual checks in 2016 is comparatively weak, with a recall of 0.76 (Table 3). Accordingly, only the datasets from 2017 to 2019 should be used for training the final model. However, the final model based on the aggregated datasets should deliberately not be allowed to assign a data point to a specific year. Otherwise, the NN would recognize a different pattern for each year and not learn holistically with all data points.

The NN based on 2017, 2018 and 2019 data detects 88 % of non-dropouts. The precision is 0.94. It means that most of our dataset is unproblematic for further use and does not need to be checked in more detail. The dropout recall is 0.76, so 76 % of dropouts are also detected as such and removed from the person dataset. In addition, the trips of the dropouts are removed from the trip dataset. For 2017, 2018 and 2019, approximately 130 participants were identified as dropouts during the pre-checks and removed (Table 2).

The performance for the aggregated dataset from 2017, 2018, and 2019 is also shown in Table 2. It is based on the personal data mentioned in Table 1.

The recall of 0.88 means that 88 % of non-dropouts are identified. At the same time, 94 % of non-dropouts are not dropouts. Furthermore, 76 % of the dropouts are identified. Thus, the model classifies 404 out of 539 individuals as non-dropouts and 135 individuals as dropouts.

Table 3 summarizes the NN's dropout identification results during the individual checks ((2), Fig. 1). Again, the results are based on 20 repetitions. While the recall for the class of dropouts in 2015 and 2016 is only 0.78 and 0.76, respectively, it is at least 0.85 in 2017, 2018, and 2019. Thus, the final model will also be trained with the aggregated datasets from these three years. The recall for the model trained with the aggregated datasets from 2017, 2018, and 2019 is 0.91, and the precision is 0.22. Thus, 91 % of the actual 196 dropouts are detected by the NN. In total, 718 participants are labeled as potential dropouts for this purpose. 8,789 participants are classified as non-dropouts. However, the precision of this model is poor (Table 3). This is due to the problem of label imbalance, which is common in travel survey data. The imbalance of the sample results mainly from the fact that there are still many more people in the MOP who report well and completely.

When looking at the label frequencies, it becomes clear that dropouts are rarely identified during the individual checks. For example 2019, out of 3,152 participants for whom the individual trip checks were performed, only 86 participants were classified as dropouts. This results in a very unbalanced class frequency, so the dropouts class must be weighted nearly 50 times as heavily as the class of non-dropouts. As a result, the model must learn the patterns of a dropout from very few data points. This is also evident from the low precision of 0.21 for 2019: the model must classify many individuals as dropouts to include the few actual dropouts.

4.2. Performance of the models for the trip checking

An overview of the performance of each model – NN, Decision Tree, Random Forest and SVM – is summarized in Table 4. The analysis is based on the trip data displayed in Table 1. The training/test ratio is again 80/20 and the *binary_crossentropy* is used as loss function. One focus in comparing the models is on the recall of class P. Recall is more important than precision when the cost of acting is low, but the opportunity cost of skipping an example is high. In this case, the cost of action represents the time the agent spends. However, the opportunity cost of classifying a class P data point incorrectly is that the data point is not checked. This in turn leads to lower data quality.

In this context, all trips that may need to be edited must also be labeled as such so that the trained staff consider them during the manual case-by-case checking afterward. Thus, it is less bad if a trip is incorrectly assigned to class P than if it is incorrectly assigned to class Ok. Accuracy, as a measure of the proportion of correctly classified data points out of the total number of data points, is also an indicator of

Table 2
Performance metrics of the NN for dropout identification during pre-checks (1) based on 20 training repetitions.

Year	Class	Accuracy	Recall	Precision	Predicted class frequency	True class frequency
2015	Dropout	0.61	0.62	0.55	72	43
	No dropout		0.54	0.78	66	95
2016	Dropout	0.74	0.69	0.88	26	54
	No dropout		0.92	0.68	118	81
2017	Dropout	0.77	0.55	0.52	40	43
	No dropout		0.86	0.84	137	134
2018	Dropout	0.71	0.73	0.46	73	44
	No dropout		0.73	0.90	117	146
2019	Dropout	0.82	0.75	0.37	33	18
	No dropout		0.87	0.95	139	154
2017	Dropout	0.84	0.76	0.62	135	105
2018	No dropout		0.88	0.94	404	434
2019						

Table 3
NN performance metrics for dropout identification during individual checks (2) based on 20 training repetitions.

Year	Class	Accuracy	Recall	Precision	Predicted class frequency	True class frequency
2015	Dropout	0.94	0.78	0.35	151	68
	No dropout		0.96	0.99	2,601	2,684
2016	Dropout	0.91	0.76	0.26	260	89
	No dropout		0.91	0.99	2,679	2,849
2017	Dropout	0.94	0.85	0.25	204	60
	No dropout		0.94	1	2,939	3,083
2018	Dropout	0.87	0.86	0.11	390	50
	No dropout		0.88	1	2,736	3,076
2019	Dropout	0.92	0.95	0.21	389	86
	No dropout		0.93	1	2,849	3,152
2017	Dropout	0.93	0.91	0.22	718	196
2018	No dropout		0.93	1	8,789	9,311
2019						

Table 4
Comparison of individual trip checking models using Accuracy, Recall, and Precision performance metrics based on 20 training repetitions; OK = no further trip checking needed, P = Trips that may need to be further checked.

		NN			Decision tree			Random Forest			SVM		
Year	Class	Accuracy	Recall	Precision	Accuracy	Recall	Precision	Accuracy	Recall	Precision	Accuracy	Recall	Precision
2015	Ok	0.89	0.90	0.98	0.92	0.95	0.97	0.93	0.95	0.97	0.86	0.98	0.87
	P		0.86	0.45		0.72	0.56		0.73	0.60		0.83	0.38
2016	Ok	0.89	0.90	0.99	0.94	0.96	0.98	0.94	0.97	0.97	0.87	0.87	0.99
	P		0.88	0.42		0.71	0.57		0.68	0.64		0.87	0.35
2017	Ok	0.81	0.87	0.92	0.77	0.81	0.93	0.75	0.78	0.93	0.80	0.85	0.92
	P		0.35	0.29		0.48	0.23		0.48	0.20		0.34	0.22
2018	Ok	0.87	0.88	0.90	0.92	0.94	0.98	0.93	0.96	0.97	0.87	0.87	0.98
	P		0.84	0.29		0.58	0.37		0.56	0.46		0.77	0.28
2019	Ok	0.84	0.84	0.98	0.89	0.92	0.96	0.90	0.94	0.96	0.83	0.84	0.98
	P		0.77	0.27		0.50	0.29		0.47	0.36		0.72	0.22
2015	Ok	0.87	0.87	0.98	0.90	0.93	0.97	0.91	0.94	0.97	0.85	0.85	0.98
2016	P		0.82	0.33		0.66	0.41		0.65	0.44		0.79	0.29
2018													
2019													

performance as an overarching metric. However, it should not be weighted too heavily due to the imbalance of class frequencies.

All four models can only classify the trips from 2017 very poorly. This can be seen in Table 4 by the low recall of class P for all four models. A more detailed examination of the dataset reveals this: Unlike in 2015, 2016, 2018, and 2019, the edited trips for 2017 cannot be assigned to the unedited trips. Therefore, the final model is trained with an aggregated dataset from 2015, 2016, 2018, 2019.

When examining the recall of class P, it is found that the NN classifies the trips best, closely followed by the SVM. The largest difference in the performance of the NN and SVM is in 2018, with the NN achieving a recall of 0.84 and the SVM achieving a recall of 0.77 this year, which is 7 % lower than the NN. This year's large difference in performance suggests that learning from the 2018 training data would require parameter adjustment in the SVM. However, the SVM can reproduce the aggregated data of 2015, 2016, 2018, and 2019 very well, with a class P recall of 0.79.

The Decision Tree and Random Forest have a lower performance than the NNs and SVM, even though their precision is sometimes higher than the precision of the NNs. In some cases, only about 50 % of the trips to be checked are classified correctly. Thus, they are not suitable for practical use. In 2017, the class P of the Decision Tree and the Random Forest recall exceeded the NN and the SVM recall. However, since the trips often do not have correct labels this year, this shows this classification's randomness.

Thus, the NN performs the best. The model trained with aggregate trip data from 2015, 2016, 2018, and 2019 achieves a class P recall of 0.82 and a precision of 0.33, meaning that 82 % of the trips to be checked are classified as belonging to class P. This is the best performance for the SVM. Of the data points assigned as class P, 33 % actually belong to this class. About 2/3 of the trips assigned to class P are incorrect in this class. This is subordinate in that it is more important to capture as many trips of class P as possible than to reduce the number of trips in class P. It is better one trip too many is checked by the editors than that a trip that should be checked for plausibility is not checked.

It is noticeable that through the NN, the trips from 2015, 2016, and 2018 can be better classified with a recall of class P above 0.84 each than in 2019 with a recall of 0.77. While many loop trips (e.g. walking the dog) are identified in 2015 and 2016, this is much less often the case in 2018 and 2019. This is because, since 2018, Kantar (fieldwork agency) has often identified these through comments made by participants on their reported trips. However, the percentage of trips that need to be inserted, merged, or deleted is much higher in 2018 and 2019, at 1.2 % and 1.39 %, respectively, than in 2015 and 2016, at 0.53 % and 0.31 %, respectively. Because inserting, merging, and deleting trips is challenging for the model to learn than identifying, for example, circular trips, the NN's performance decreased slightly in 2019.

4.3. *Effort reduction in terms of trips and participants to be checked*

The reduction in effort in absolute terms for the number of trips to be checked is the added value of this study for further use. It is shown in Fig. 2 and refers to the aggregated data for 2015, 2016, 2018 and 2019. Of the 274,273 total trips, 49,253 are assigned to class P and must be checked manually. The exact number of trips may differ slightly depending on the training run of the model and the random split into training and test data. Thus, the staff must check only 17.7 % of the trips.

In addition, the trips can be distinguished in terms of the check's difficulty. If only the start or end time has to be changed, this is relatively easy for the staff to recognize. This is the case for 7,547 of the 49,253

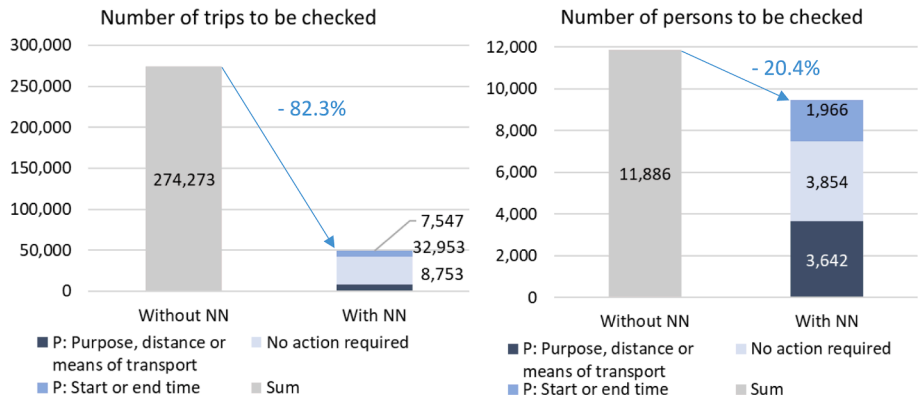


Fig. 2. Effort reduction in terms of persons and trips to be checked for 2015, 2016, 2018, and 2019 (aggregated).

trips. If the trip purpose, the transport mode, or the distance of a trip needs to be changed, it may be necessary to also look at a participant's trip history. Furthermore, it makes the validation more time-consuming for the staff. This is the case for 8,753 trips. This step is also needed if the trip is incorrectly assigned to class P. This is the case for 32,953 trips. After all, this is the only way to verify if the trip purpose, distance and means of transport are plausible.

Accordingly, it is important to consider the reduction in the number of trips and examine the number of persons to be checked. The reduction in effort for this is summarized in Fig. 2. The 49,253 trips to be checked come from 9,462 different persons. Since there were 11,886 participants in 2015, 2016, 2018, and 2019, the effort reduction here is only 20.4 %. That means that 79.9 % of the participants still have to be checked. However, for 1,966 of the 9,462 participants, only one trip's start or end time has to be checked. Thus, only 6,454 participants need to be checked for their trip history to determine whether the trip purpose, the distance, or the transport mode needs to be adjusted or if no change is necessary. This reduces by 36.9 % the number of participants for whom the complete trip history must be considered time-consuming.

5. Discussion and conclusion

Machine learning has met with varying degrees of success in dropout identification and individual trip checks. It was shown that machine learning can achieve a more efficient performance compared to manual checks. However, no model that allows fully automated data checks without human assistance could be found and implemented. The models presented can only provide a recommendation for action, which can significantly reduce the workload when considered as a whole. In the following, we look at where using machine learning models could be beneficial.

5.1. Dropout identification during pre- and individual checks

For some participants, it is unclear if they may have stopped reporting. Such 'abnormal' persons could report authentic travel patterns without being directly removed from the datasets. Therefore, these 'abnormals' must be given special attention. This means the classification into dropouts and non-dropouts cannot be made unambiguously, even by the trained staff. However, the performance of the NN can be seen as very positive. Since the NN can fully take over the process of dropout identification during the pre-checks, a fully automated solution approach can be presented. Furthermore, the NN can enable a consistent classification. Normally, borderline cases are always assigned to the same class by the trained staff. The application of an NN for dropout identification during pre-screening is thus possible since the staff's decisions regarding the classification can be learned effectively.

For dropout identification during individual checks, it is essential to identify all possible dropouts so that the staff is aware of them during editing. With a recall of 0.91 for the class of dropouts based on the 2017, 2018 and 2019 data, NN works well. However, the precision is relatively low at 0.22. This is also evident in the absolute numbers of the classes. Within three years, there were 196 actual dropouts. However, 718 individuals received the label that they might be dropouts (Table 3). This means that the model has to classify many participants as dropouts to identify the few actual dropouts. With 9,311 participants who are not dropouts, there are only 196 dropouts (unbalanced classes). This leaves the model with few data points to learn a pattern for dropouts.

For the individuals who are misclassified, it can be assumed that the classification of these individuals is not entirely clear. Even for the trained staff, which form the learning basis for the classification, it is not entirely clear in some cases which participant has stopped reporting or whether the few or unreported trips correspond to the participant's actual travel behavior.

Using the NN leads to a significant reduction in the effort since instead of the original 9,507 persons within three reporting years, only 718 would have to be considered. However, even without using the NN, it is obvious that some of the 9,507 persons are not dropouts. This can be seen from many reported trips spread throughout the week. Based on the data, it is not quantifiable how much effort can be saved in absolute terms due to missing information in the data and data privacy reasons. In any case, using the NN can serve as decision support.

5.2. Individual trip checks

The application of machine learning results in a binary classification. One class is for trips that do not need to be edited because they were correctly reported. The second class is for trips that may need further editing and should be checked accordingly by trained staff. The classification is constructed to reduce the workload for the trained staff. Independent editing by the model is impossible because single edits occur only rarely, so the model cannot learn this.

We applied four different machine learning models. A comparison of the models shows that the NN is the most effective model compared to a Decision Tree, Random Forest, and SVM. The application of the NN leads to a considerable reduction in the number of trips to be checked.

5.3. General conclusion on the research framework

This paper investigates the applicability of machine learning for checking the quality of participant responses and their reported trips in the MOP. Due to the long history of the MOP and many years of manual data checking by trained staff, data is provided to support the model's training. With the implementation of the new approach to check trips and participants using machine learning techniques, the checking effort, which in the end still has to be taken over by trained staff, can be reduced. The presented approach follows the maxim that as much as necessary but as little as possible is changed in the data. However, the

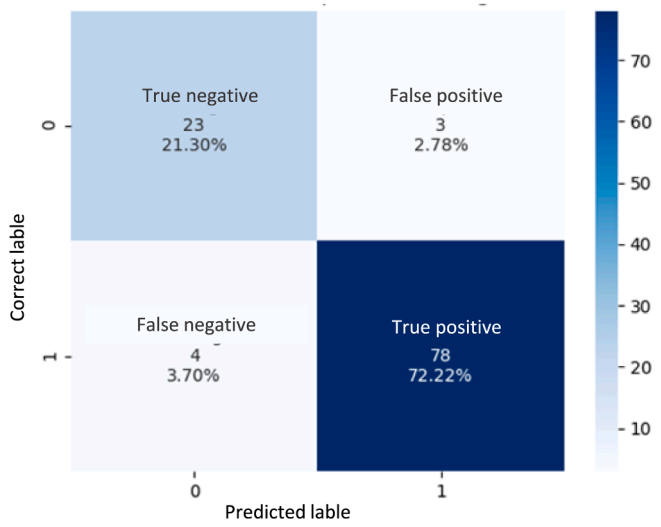


Fig. 3. Confusion matrix for dropout identification during pre-check.

model performance is comparatively poor, so it must be emphasized that the trade-off between precision and the number of cases to be checked by staff must always be considered when implementing such an approach. In the future, techniques to mitigate overfitting, such as incorporating dropout layers, early stopping, or skip connections, can be integrated to improve the models.

In the field of survey research, such approaches to data checking as presented in this paper are still in the early stages of research, especially in survey methodology and social science. The literature review shows that such approaches have been used primarily with technically collected data. Since it was shown in this study that machine learning can help to eliminate the need to individually verify one in five individuals, it is expected that this methodology promises significant time and personnel savings and cost savings. The presented methods have a great potential to make research processes cost-effective in the future because a large part of the costs in the data processing process are related to data checking.

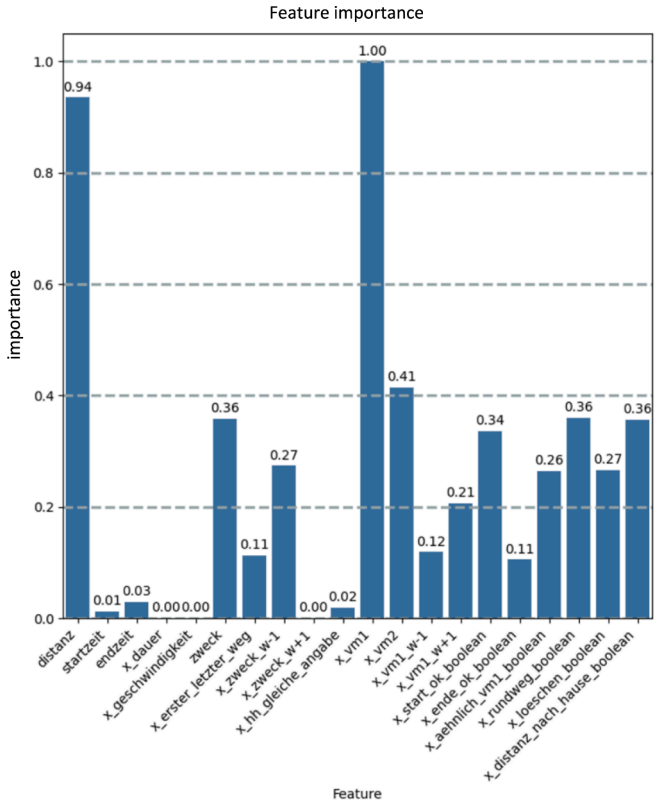


Fig. 4. Feature importance of all features for the NN

5.4. Shortcomings

Although our study was able to show that the data validation of trips can be taken over by machine learning, our study has shortcomings: Because the MOP has a consistent framework in terms of, e.g., survey method and questionnaire design for years, we use data from the past as ground truth data to learn. Suppose new modes of transportation (e.g., e-scooters) are steadily becoming more popular and used. In that case, they are underrepresented in the data, or the algorithm does not know how to handle them. The reasons for this are the small number of cases and the low variance. This is dangerous for the outcomes because the goal is not to create "perfect" trip diaries but to guarantee content consistency and completeness. As discussed in Aschauer et al. (2018), the general question of survey data bias is used as "ground truth" data for comparison.

5.5. Outlook

This work presents new approaches to checking travel survey data of the MOP, in which data at the individual level (sociodemographic characteristics of a person and trips made) were checked. The main advantage of this work is that already checked data from the past is available to let an algorithm learn checking routines of the past, thus replicating the approach of trained staff.

In response to how machine learning approaches can help check trip data from the MOP, several validation and measures were identified and elaborated, which are also found in the training data. It was shown that the presented approach can reduce but not completely replace the checking processes by the trained staff in terms of effort. This is because of the lack of information and the high inter- and intrapersonal variance, e.g., in movement patterns over a week.

In this exploratory study, the algorithms were applied to the MOP data only. The question arises to what extent the presented algorithms will apply to other studies in the future. In assessing the extent to which the algorithms are transferable, it must be considered that conditioning effects may occur in the MOP data (e.g., omission/summation of very short trips). Since the response burden in the MOP is comparatively high compared to cross-sectional surveys, it cannot be quantified how these method artifacts affect the algorithms and if the algorithms can be applied on other survey data. Another issue to be considered for transferability is the ground truth data. This study used data from previous years of the MOP to check the new dataset. For German studies, a validation based on these data might be possible. However, it needs to be checked whether the MOP data can be used as ground truth data for surveys from other countries or if additional data are available for learning so that effects resulting from method artifacts can be minimized. Furthermore, it is important to note that, to the authors' knowledge, there is no worldwide standard for data preparation of NHTS. This means that data management may differ from survey to survey. For example, in the MOP, excluding individuals from the dataset is allowed. If the framework were to be applied to another dataset, it would have to be checked whether all three steps presented are necessary for the study or only individual steps (e.g., trip validation) should be adopted. However, the presented framework is designed so that single modules can be extracted and used as a single module. In conclusion, it should be emphasized that the applicability and transferability to other studies is limited. Basically, this work can be seen as a succesfull feasibility study that can be used as a stimulus to use Machine Learning applications for data checking in other contexts.

The results presented in this paper help improve and harmonize data checking in the MOP and provide insights into the limitations of the

methods presented. Overall, it can be concluded that machine learning methods can offer support for checking travel behavior data.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

CRediT authorship contribution statement

Lisa Ecke: Conceptualization, Methodology, Writing – original draft, Validation. **Miriam Magdolen:** Conceptualization, Writing – original draft, Validation. **Sina Jaquart:** Methodology, Software, Data curation, Visualization, Investigation, Validation. **Robin Andre:** Methodology, Validation. **Peter Vortisch:** Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Appendix

Table 5
Overview of the features for training the dropout identification during the pre-check.

Feature	Value	Characteristics
Age	Stetig	{1,2,…, 97}
Mobility restriction	categorical	yes no
Job	categorical	Unemployed In education In vocational training Retired Not employed, housewife Part-time employed Fully employed
Householdtype	categorical	Household with children under 18 Household without children, 3 and more adults Small household with employed persons (1–2 persons)
Number of reported days	continuous	{1, 2,…, 7}
Number of reported trips	continuous	{1, 2,…, 51}
Number of days with illness	continuous	{1, 2,…, 7}
Number of days on vacation	continuous	{1, 2,…, 7}
Number of days on which abnormalities were reported	continuous	{1, 2,…, 7}

Table 6
Overview of the features used to train the final model for the individual trip checks.

Feature	Description	Construction	Characteristics
distance	Reported distance of the trip	–	(0, 1000]
Trip purpose	Reported purpose of the trip	–	Way to work Official/business route Way to the educational institution errand, shopping

(*continued on next page*)

Table 6 (*continued*)

Feature	Description	Construction	Characteristics
			Leisure route Service route Home Other Way to out of home Way to 2nd residence Other private errand Loop trip
x_erster_letzter_weg	Classification, which is the trip of the day	Check whether the trip took place directly before or after a day change	First trip Last trip In between
x_zweck_w-1	Purpose of the trip before on the same day	Consideration of the day changes and the trip sequence	see feature trip purpose
x_vm1	Determination of the main means of transport used to travel the trip	Based on the mode of transport ranking: airplane > long-distance public transport > local public transport > MIV driver > MIV passenger > bicycle/pedelec > foot. The mode of transportation used with the highest rank is determined to be the primary mode of transportation.	On foot Normal bike, electric bike& pedelec Moped, motorcycle Car as driver Car as passenger City bus, regional bus Long-distance bus, coach streetcar, light rail, subway suburban train, regional train Long-distance train Airplane Other Ship Truck Horse, carriage Motorhome
x_vm2	Determination of a second main means of transport by which the trip was travelled	Cf. x_vm1 for the case when more than one means of transport was used for the trip	See x_vm1
x_vm1_w + 1	Main means of transport of the way afterwards on the same day	x_vm1 is taken over from the trip afterwards, if it took place on the same day	see x_vm1
x_start_ok_boolean	Rule-based guess whether the start time was specified correctly	Verify that the trip should not start, for example, before the trip has previously ended	Yes No
x_ende_ok_boolean	Rule-based guess whether the end time was specified correctly	Checking possible cases. For example, whether there is at least one minute difference between the end time of the trip and the start time of the following trip or whether the end time of the last trip is after the end of the last reporting day.	Yes No
x_aehnlich_vm1_boolean	Verify that similar trips are reported by a person using the same primary mode of transportation	First, similar trips must be identified within the person's trip history. Then, it is checked whether the reported main modes of transport match.	Yes No
x_rundweg_boolean	Checking whether it could be a loop trip	Certain means of transport, such as "on foot" or "inline skating", in conjunction with corresponding successive trip purposes, such as "leisure trip" or "other private errand", indicate a circular route.	Yes No
x_loeschen_boolean	Rule-based check whether the trip should be deleted	If, for example, leisure trips are combined into a loop trip, business trips and trips to work are linked or intermodal trips can be combined, individual trips must be deleted.	Yes No
x_distanz_nach_hause_boolean	Checking the distance of the trip back home	The distance of the trip home should be in a certain range around the distance how far the person has gone from home before.	Yes No

References

Armoogum, J. (Ed.), 2014. Survey Harmonisation with New Technologies Improvement, SHANTI. IFSTTAR; COST, Bruxelles.

Aschauer, F., Hössinger, R., Axhausen, K.W., Schmid, B., Gerike, R., 2018. Implications of survey methods on travel and non-travel activities: A comparison of the Austrian national travel survey and an innovative mobility-activity-expenditure diary (MAED). Eur. J. Transp. Infrastruct. Res. 18 (1), 2018. https://doi.org/10.18757/ejtir.2018.18.1.3217.

Aschauer, F., Hössinger, R., Jara-Diaz, S., Schmid, B., Axhausen, K., Gerike, R., 2021. Comprehensive data validation of a combined weekly time use and travel survey. Transp. Res. Part A: Policy Pract. 153, 66–82. https://doi.org/10.1016/j.tra.2021.08.011.

Bonnel, P., Bayart, C., Smith, B., 2015. Workshop synthesis - Comparing and combining survey modes. Transp. Res. Procedia 11, 108–117. https://doi.org/10.1016/j.trpro.2015.12.010.

Chlond, B., Wirtz, M., Zumkeller, D., 2013. Do dropouts really hurt? – Considerations about data quality and completeness in combined multiday and panel surveys. In: Zmud, J.P., Lee-Gosselin, M., Munizaga, M.A., Carrasco, J.A. (Eds.), Transport Survey Methods - Best Practice for Decision Making. Emerald, Bingley.

Couper, M.P., 2011. The future of modes of data collection. Public Opin. Q. 75, 889–908. https://doi.org/10.1093/poq/nfr046.

Haas, M. de, Kroesen, M., Chorus, C., Hoogendoorn-Lanser, S., Hoogendoorn, S. (Eds.), 2022. Didn't Travel or Just Being Lazy? An Empirical Study of Soft-Refusal in Mobility Diaries.

Dreyfus, G., 2005. Neural Networks: Methodology and Applications. Springer, Berlin Heidelberg, Berlin, Heidelberg.

Ecke, L., Chlond, B., Magdolen, M., Vallée, J., Vortisch, P., 2021. Deutsches mobilitätspanel (MOP). In: – Wissenschaftliche Begleitung Und Auswertungen Bericht 2020/2021. Alltagsmobilität Und Fahrleistung, p. 152 pp.

Ecke, L., Vallée, J., Chlond, B., 2022. German Mobility Panel - Longitudinal study on the travel behavior of the population: Project website. Karlsruher Institut für Technologie. https://mobilitaetspanel.ifv.kit.edu/english/index.php (accessed 30 September 2022).

Ecke, L., Hilgert, T., Magdolen, M., Chlond, B., Vortisch, P., 2024. Checking data quality of longitudinal household travel survey data. Transp. Res. Proc. 76, 258–268. https://doi.org/10.1016/j.trpro.2023.12.053.

Feng, T., Timmermans, H.J., 2013. Transportation mode recognition using GPS and accelerometer data. Transp. Res. Part C: Emerg. Technol. 37, 118–130. https://doi.org/10.1016/j.trc.2013.09.014.

Hu, S., Lin, H., Chen, X., Xie, K., Shan, X., 2022. Modeling usage frequencies and vehicle preferences in a large-scale electric vehicle sharing system. IEEE Intell. Transport. Syst. Mag. 14, 74–86. https://doi.org/10.1109/MITS.2019.2953561.

Hu, S., Xiong, C., 2023. High-dimensional population inflow time series forecasting via an interpretable hierarchical transformer. Transp. Res. Part C: Emerg. Technol. 146, 103962 https://doi.org/10.1016/j.trc.2022.103962.

Hubrich, S., Wittwer, R., Gerike, R., 2018. Quality indicator set for household travel surveys. Transp. Res. Procedia 33, 219–226. https://doi.org/10.1016/j.trpro.2018.10.098.

Jödden, C., Führer, M., 2021. Deutsches Mobilitätspanel (MOP) – Erhebung der Alltagsmobilität sowie der Pkw-Fahrleistungen und Kraftstoffverbräuche: Endbericht zum Paneljahr 2020/2021. Kantar, München, 141 pp. (accessed 17 August 2022).

Kitamura, R., Bovy, P.H., 1987. Analysis of attrition biases and trip reporting errors for panel data. Transp. Res. Part A: Gen. 21, 287–302. https://doi.org/10.1016/0191-2607(87)90051-3.

Kralj Novak, P., Šmuc, T., Džeroski, S. (Eds.), 2019. Variance-Based Feature Importance in Neural Networks. Springer International Publishing; Imprint Springer, pp. 306–315.

Kuhnimhof, T., Chlond, B., Zumkeller, D., 2006. Nonresponse, selectivity, and data quality in travel surveys: Experiences from analyzing recruitment for the german mobility panel. Transp. Res. Rec.: J. Transp. Res. Board 29–37.

Lundberg, S., Lee, S.-I., 2017. A Unified Approach to Interpreting Model Predictions.

Madre, J.-L., Axhausen, K.W., Brög, W., 2007. Immobility in travel diary surveys. Transportation 34, 107–128. https://doi.org/10.1007/s11116-006-9105-5.

Meng, C., Cui, Y., He, Q., Su, L., Gao, J., 2017. Travel purpose inference with GPS trajectories, POIs, and geo-tagged social media data. IEEE Int. Conf. Big Data 2017, 1319–1324. https://doi.org/10.1109/BigData.2017.8258062.

Montini, L., Rieser-Schüssler, N., Horni, A., Axhausen, K.W., 2014. Trip purpose identification from GPS Tracks. Transp. Res. Rec. 2405, 16–23. https://doi.org/10.3141/2405-03.

NHTS, 2018. 2017 NHTS Data User Guide. Federal Highway Administration Office of Policy Information, p. 76. https://nhts.ornl.gov/assets/NHTS2017_UsersGuide_04232019_1.pdf.

von Behren, S., Hilgert, T., Kirchner, S., Chlond, B., Vortisch, P., 2020. Image-based activity pattern segmentation using longitudinal data of the german mobility panel. Transp. Res. Interdisc. Perspect. 8, 100264 https://doi.org/10.1016/j.trip.2020.100264.

Wang, B., Gao, L., Juan, Z., 2018. Travel mode detection using GPS data and socioeconomic attributes based on a random forest classifier. IEEE Trans. Intell. Transport. Syst. 19, 1547–1558. https://doi.org/10.1109/TITS.2017.2723523.

Wang, T., Hu, S., Jiang, Y., 2021. Predicting shared-car use and examining nonlinear effects using gradient boosting regression trees. Int. J. Sustain. Transp. 15, 893–907. https://doi.org/10.1080/15568318.2020.1827316.

Wirtz, M., Streit, T., Chlond, B., Vortisch, P., 2013. On new measures for detection of data quality risks in mobility panel surveys. Transp. Res. Rec.: J. Transp. Res. Board 19–28. https://doi.org/10.3141/2354-03.

Wittwer, R., Hubrich, S., 2015. Nonresponse in household surveys: a survey of nonrespondents from the repeated cross sectional study "mobility in cities – SrV" in Germany. Transp. Res. Procedia 11, 66–84. https://doi.org/10.1016/j.trpro.2015.12.007.

Zhou, X., Yu, W., Sullivan, W.C., 2016. Making pervasive sensing possible: Effective travel mode sensing based on smartphones. Comput. Environ. Urban Syst. 58, 52–59. https://doi.org/10.1016/j.compenvurbsys.2016.03.001.