Available online at www.sciencedirect.com

ScienceDirect

Transportation Research Procedia 00 (2019) 000–000

Transportation Research Procedia

www.elsevier.com/locate/procedia

International Steering Committee for Transport Survey Conferences

Transferability analysis of user groups in travel behaviour surveys using a random forest classification model

Simon Nieland[a], Rebekka Oostendorp[a], Matthias Heinrichs[a], Rita Cyganski[a]

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

Abstract

There are many different travel behavior surveys with a wide range of sample sizes and contents, whose data sets are often difficult to merge. This contribution aims to evaluate the possibilities of transferring user typologies from one travel survey to another based on two different surveys in Berlin, Germany, using a random forest classification model. The investigated unimodal and intermodal mobility types were generated on the basis of a travel survey (n=1,098), collected in the year 2016 in Berlin with a special focus on intermodality, and were transferred into the Germany-wide survey "Mobility in Germany" (MiD) from 2017 (n=316,361 (total sample Germany); n=3,206 (subsample Berlin)). Basis for the training of the random forest model were mobility resources (e.g. public transport ticket availability, number of cars in household), socio-demographic characteristics (e.g. size of household, age, employment), temporally aggregated uni- and intermodal usage frequencies and trip purposes. At first, the model has been developed and tested based on the Berlin survey using different subsets of input variables (e.g. without usage frequencies, with usage frequencies, without intermodal usage) to evaluate which classification accuracy can be achieved depending on what kind of variables are included in the survey. In this process, gradual reduction of the variables was performed to evaluate the effects of using a reduced number of input variables for transfer. Based on usage frequencies, socio-demographics and mobility resources, the classification achieved a mean F1 score of 0.93 for the mobility types in the Berlin survey. Respectively, the results were lower when reducing the number of training variables. When performing the transfer, the distribution of the resulting user groups shows high similarities, especially for the Berlin sample. In conclusion, it can be shown that the proposed methodological procedure is suitable for transferring mobility types developed on the basis of a specialized data set to other surveys on the basis of mobility behavior parameters.

Keywords: Random forest classification; Mobility types; Data fusion; Machine learning

1. Introduction

There is a variety of travel behaviour surveys with a wide range of thematic information, whose data sets are often difficult to merge. The possibility of extrapolating information from one survey to another comprises huge potentials in terms of fostering applicability of existing surveys by adding additional information. Data fusion of multiple surveys with different origins has been discussed in the literature for some time Bayart et al. (2009); Goulias (2000). Besides record-linkage, which can be used when single records can be linked from one survey to another, the k-Nearest Neighbor (k-NN) method is regularly used for continuous data and the multivariate imputation by chained equations (MICE, Van Buuren and Oudshoorn (1999)) for mixed data types. Most prevalent methods for categorical survey data fusion are multinominal logistic regression models Bahamonde-Birke and Hanappi (2016); Heldt et al. (2018). These parametric models depend on the selection of suitable parameters, include assumptions about the influences of the parameters on the destination variable and make assumptions about the distribution of the data and its variables Stekhoven and Bühlmann (2012). Random Forest (RF) Breiman (2001) is a non-parametric method for classification and regression problems which does not make assumptions about the distribution of the variables, can deal with any kind of data types (categorical, continuous and mixed-types) and is able to solve regression/classification problems with strongly non-linear relationships. It has been successfully applied to data imputation problems in the domain of bioinformatics and medical informatics Shah et al. (2014); Stekhoven and Bühlmann (2012). Random forests regularly outperform classical methods like k-NN and MICE and do not require prior domain knowledge Shah et al. (2014); Stekhoven and Bühlmann (2012); Waljee et al. (2013). For this reason, this study uses the Random Forest classifier to transfer the affiliation of a single person to a certain class of a user typology from a local, highly specialized survey to the national travel survey of Germany.

This typology of mobility behaviour was generated using a clustering approach based on a local travel survey with a focus on intermodal travel behaviour Oostendorp et al. (2019). Mobility user typologies represent an effective possibility to simplify the highly diverse mobility behaviour of individuals, thus making the typologies more applicable and suitable for practitioners and subsequent analyses Oostendorp et al. (2019). Typologies are commonly generated using segmentation approaches to categorize individuals with a certain travel behaviour which can then be used to develop user-specific applications. Today, segmentation approaches are an established means of analysing daily travel determinants and are used by different disciplines such as psychology Hunecke et al. (2010), sociology Jensen (1999), and increasingly also by transport sciences Hildebrand (2003); Krizek and Waddell (2002), Krizek and Waddell (2002); Outwater et al. (2004); Prillwitz and Barr (2011). User typologies are normally generated on the basis of travel surveys using grouping procedures and therefore represent a certain sample of a specific part of the population. The advantage of such grouping approaches is that cause-effect relationships are in certain cases only detectable in subgroups of the total population, depending on the specific research focus. Furthermore, grouping allows better communication between scientists and practitioners by identifying homogeneous groups so as to reduce the complexity of heterogeneous populations Haustein and Hunecke (2013). In order to generate universally applicable user groups, transferability of the identified user groups into other surveys and thus the generation of additional value in existing data sets is desirable. This is especially important since approaches can be developed in surveys with small sample sizes and then transferred to bigger surveys based on a certain set of attributes.

Accordingly, the main objective of this study is to evaluate the possibilities of transferring user typologies from a small travel survey of Berlin to a country-wide survey using a random forest classification model.

2. Methodology

2.1. Data

The local, highly specialized survey investigated unimodal and intermodal mobility types of the reporting persons. The travel survey (n=1,098) was collected in the year 2016 in Berlin with a special focus on intermodality Oostendorp and Gebhardt (2018) . By definition, intermodal mobility behaviour means that people use and combine at least two different modes of transportation on a single trip Chlond (2013); Jones et al. (2000). For the generation of the mobility types, information on mobility behaviour (frequency of use of specific means of transport and trip purposes) was used as input Oostendorp et al. (2019). The mobility types were transferred to the Germany-wide national travel survey

‘Mobility in Germany’ (MiD) from 2017 (n=316,361 total sample Germany; n=3,206 subsample Berlin) Nobis and Kuhnimhof (2018). In addition to the specific case of Berlin, large cities can be identified in the sample as a category based on population size. Basis for the training of the RF-model were mobility resources (e.g. public transport ticket availability, number of cars in household), socio-demographic characteristics (e.g. size of household, age, employment) and temporally aggregated uni- and intermodal usage frequencies of mode of transport. At first, the model was developed and tested based on the Berlin survey using different subsets of input variables (e.g. without usage frequencies, with usage frequencies, without intermodal usage) to evaluate which classification accuracy can be achieved depending on what kind of variables are included in the survey. However, the frequencies of trip purposes for each person are not available in the MiD, because it records a single day and not a longer period of time. In this process, gradual reduction of the variables was performed to evaluate the effects of using a reduced number of input variables for transfer.

2.1.1. Description of distribution of parameters in both surveys

In order to better assess the possibilities and limits of transferability, the two data sets are first compared concerning the general usage of the most important modes of transport (car, bike, public transport (PT), carsharing (CS)). In contrast to MiD, the local survey gathers intermodal and unimodal transport usage separately. Hence figures must be summarized again for a comparison. Overall, the frequency of transport use of the modes public transport, bicycle, car and car sharing in the local data set shows a plausible distribution compared to the MiD data set. Public transport is used most frequently on a daily basis, followed by bicycle and car. The frequency of use of car sharing is very low.

However, a direct comparison of the two data sets is difficult due to the different methodology applied for the reporting of transport mode usage. A systematic methodological bias seems to exist due to the separate recording of monomodal transport-mode usage and intermodal combinations and the subsequent summation of the values. This is reflected in the fact that the lower frequency categories (‘1-3 times a week’ and ‘less than monthly’) are consistently significantly lower in the local data set than in the MiD data set for both subsamples, Berlin and large cities with more than 500,000 inhabitants (Figure 1). For public transportation and bicycle usage, they are shifted towards the ‘daily use’ category, which has significantly higher values in the local data set (daily PT use: local data set 42.6 % vs. MiD data set 35.0 % (Berlin) and 30.3 % (large cities); daily bike use: local data set 31.7 % vs. MiD data set 23.2 % (Berlin) and 22.8 % (large cities)). However, the proportion of people who never use public transport or bicycles is well matched. This can be explained by the fact that intermodal combinations in most cases include a route stage covered by public transport and thus the frequency of use increases when the values are added up. As a result, the frequency of use turns out to be significantly higher with this methodological approach than with a general query of transport-mode use per transport mode (without differentiation of intermodal use) as in the MiD data set. In contrast, for car usage, the proportion of daily use in the local data set (20.5 %) maps very well to the value from the Berlin subsample of the MiD data set (21.0 %). Instead, for car and car-sharing usage, the category ‘never’ is much more prevalent in the local data set (car: 48.1 %; CS: 95.0 %) than in the MiD data set (car: 29.2 % (Berlin) and 24.4 % (large cities); CS: 88.7 % (Berlin) and 89.2 % (large cities)). Intermodal combinations in which a route stage is covered by car are rather rare overall. They are therefore less significant when adding up the frequency of use, so that the influence of monomodal usage on the total value is higher in this case.

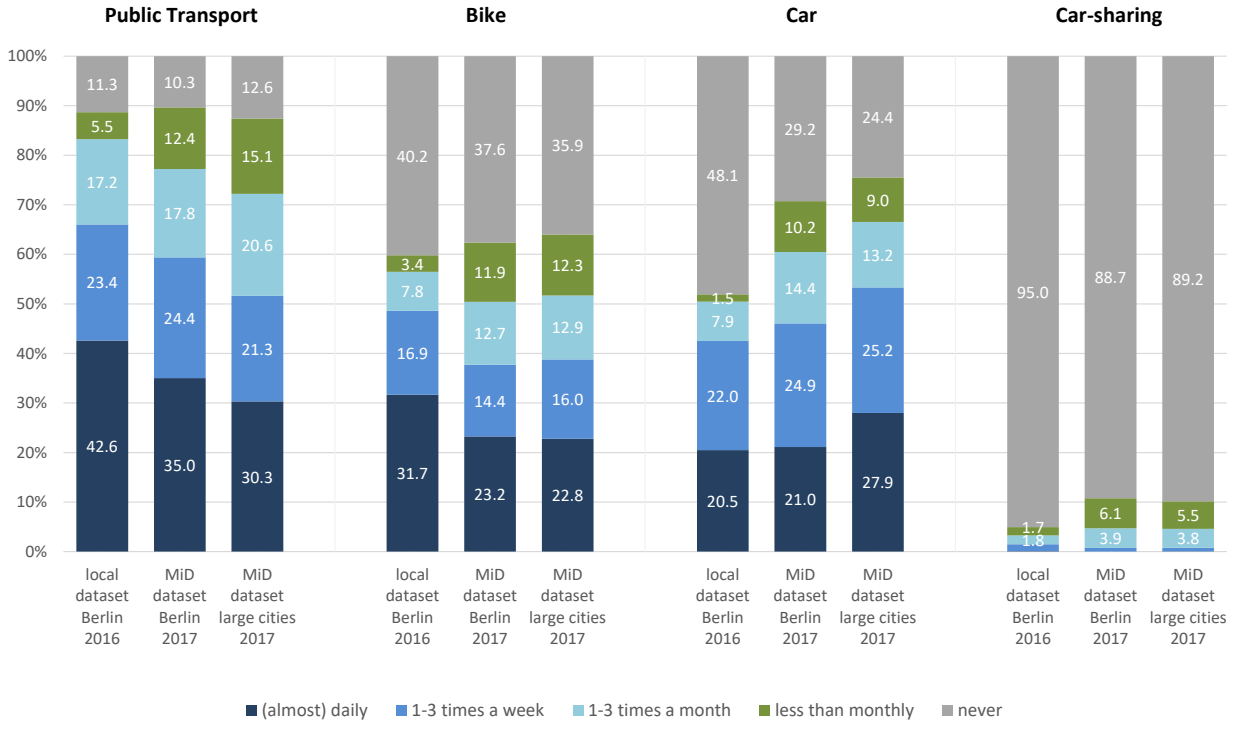

Figure 1: Frequency of transport use in the local data set (2016, Berlin, n = 1,098) and in the MiD data set (2017, subsample Berlin, n = 2,458 and subsample large cities with more than 500,000 inhabitants, n = 41,644; weighted; number of cases at participant level)

In addition to the frequency of transport mode use, the trip purposes on uni- and intermodal trips were also included as input variables in the cluster analysis to identify mobility types. Therefore, the following section shows the overall representation of trip purposes in the two data sets considered (Figure 2). Again, due to the different survey design, a different approach to the analysis was necessary. In the local data set, the frequency of mode use by trip purpose is the basis for calculating trip purposes. The frequencies for each trip purpose were summed across all modes of transportation and set in relation to each other, resulting in a relative distribution for the entire data set. For the MiD data set (sub-sample Berlin and subsample large cities > 500,000 inhabitants), the ratio of trip purposes is based on a simple frequency evaluation of the reference date data which is recorded at one single day per person and not a longer period of time.

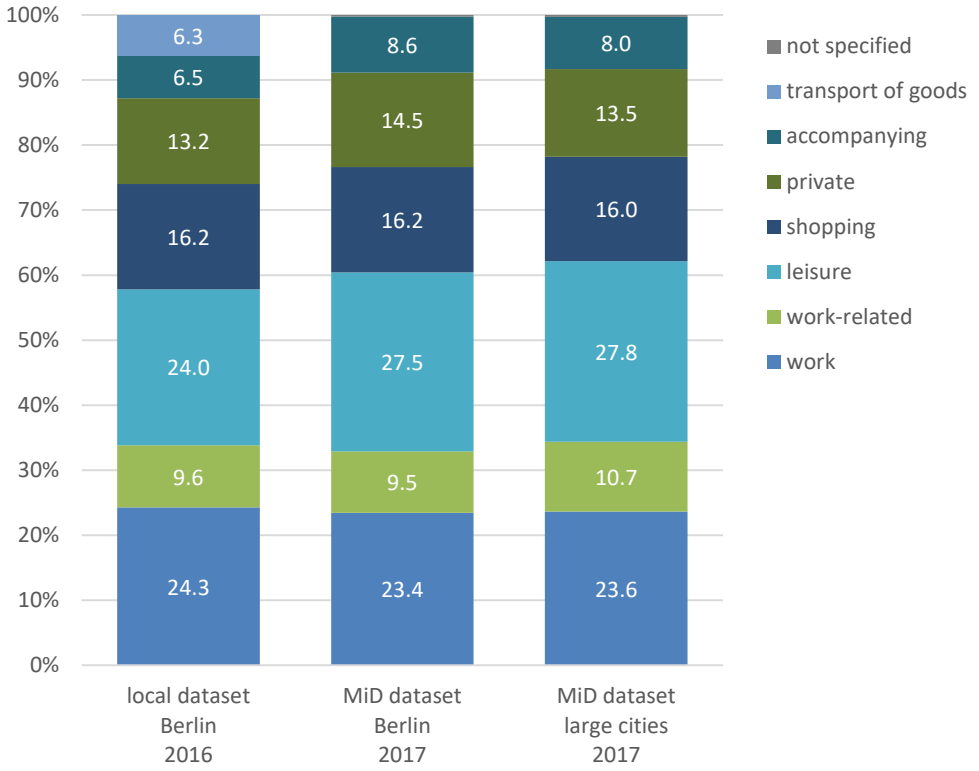

Figure 2: Distribution of trip purposes in the local data set (2016, Berlin, n = 1,098) and in the MiD data set (2017, subsample Berlin, n = 9,370 trips and subsample large cities with more than 500,000 inhabitants, n = 168,336 trips; weighted; number of cases at level of reported trips)

The distribution of trip purposes is almost the same in the local data set and in the MiD data set (both MiD subsample Berlin as well as subsample large cities with more than 500,000 inhabitants). The trip purposes leisure, picking up or taking people somewhere and private errands are slightly more represented in the MiD data set (subsample Berlin) than in the local data set (leisure: 27.5 % vs. 24.0 %; private errands: 14.5 % vs. 13.2 %; accompanying: 8.6 % vs. 6.5 %). Instead, in the local data set, the trip purpose transport of goods, merchandise, material, which accounts for 6.3% of the trip purposes, was additionally queried. This type of trip was probably assigned by respondents in the MiD 2017 to the trip purposes leisure and private errands, which explains the slightly higher expression in that case. The almost identical distribution of trip purposes in the data sets considered is a very good starting point for the transferability of the mobility types. Deviations in the transferability of the mobility types due to different trip purposes are therefore not to be expected.

Socio-demographic characteristics were not used as input variables of the mobility typing with the local data set. However, the mobility types were subsequently characterized in detail and descriptively by their predominant socio-demographic characteristics. Therefore, a comparison of the socio-demographic characteristics of the local data set and the MiD data set seems to be useful as well in order to better evaluate the success of the transferability of the types as for identifying possible deviations due to the composition of the sample later on. To enable better comparability, only the Berlin sample and only cases to which a unimodal local mobility type could be assigned were used in the evaluations within the MiD data set. The evaluation was performed unweighted to avoid additional bias as occupation, graduation, gender and age group serve as weighting factors in the MiD data set Eggs et al. (2018).

The key findings comparing the socio-demographic parameters of the two data sets can be summarized as follows:

- Age: Younger age groups are more strongly represented in the local data set than in the MiD. Accordingly, the average age in the local data set (47.77 years) is also lower than in the MiD (53.24 years).
- Gender: The gender ratio is nearly identical in both data sets (local data set: 48.4 % women to 51.6 % men; MiD: 48.5 % women to 51.5 % men).

- Income: The lower income classes are slightly more strongly represented in the local survey than in MiD, and the higher income classes are correspondingly slightly less strongly represented.
- Household size: The average household size is slightly higher in the local survey (2.25 persons) than in the MiD (2.11 persons). This is due in particular to a higher share of 3-person households and 4-and-more-person households and a correspondingly lower share of 2-person households. The share of 1-person households is about the same in both data sets.
- Occupation: The proportion of employed persons is not directly comparable due to differences in sampling (different minimum ages of participants). However, the analyses indicate that the share of employed persons is slightly higher in the local data set than in MiD.
- Mobility resources:
    - Cars: Households without own car are more strongly represented in the local data set (37.1 %) than in MiD (30.4 %). The higher car ownership in MiD is reflected in both a higher proportion of households with one car and with two cars.
    - Public transport: In the case of the availability of public transport season tickets, a direct comparison of the percentages is not possible due to different question wording in the data sets. Nevertheless, the analyses show that the participants of the local survey have a higher overall availability of public transport season tickets than the participants of MiD.
    - Car sharing: The proportion of people with car sharing membership is higher in the local survey (17.0 %) than in the MiD (12.7 %). This is consistent with the different age distribution.

    Mobility restriction: The proportion of people with a mobility restriction is lower in the local survey (6.4 %) than in MiD (13.2 %). This is consistent with the different age distribution.

2.1.2. Short description of mobility types

As presented in the introduction (section 1), the analyses in this study draw on previous work on the typology of mobility behaviour Oostendorp et al. (2019).The mobility types identified in the local Berlin survey can be described by the input variables of the cluster analysis, namely frequency of use of specific means of transport and trip purposes. They are named according to their main characteristics: public transport users (PTU), bicycle-centred users (BCU), lower activity car users (LACU) and multimodal users for all purposes (MUP). Table 1 gives an overview of the mobility types with their main contributing variables and their share in the local data set. Car users with an overall low frequency of trips represent the highest share in the dataset (LACU: 34.9 %). About one quarter of the sample can be described as everyday cyclists (BCU: 24.2 %). In this group, occasional use of car-sharing is above average. However, it should be noted that this does not mean that every person belonging to this group uses car-sharing. The public transport users (PTU) who especially use public transport unimodally for trips to work and the group of highly mobile multimodal users with many different trip purposes (MUP) account for 10.0 % each.

Table 1: Overview of mobility types and main contributing variables

Mobility type	Description	Contributing variables	Share
Lower activity car users (LACU)	car users with overall low frequency of trips	moderately positive: trips with household car strongly negative: trips by bike and by pt, all trip purposes	34.9 %
Public transport users (PTU)	unimodal commuters by PT	strongly positive: unimodal trips with pt moderately positive: work purpose strongly negative: household car usage, bike usage moderately negative: transportation of goods, shopping, car-sharing trips	20.0 %
Bicycle centered users (BCU)	everyday cyclists with occasional use of car-sharing	strongly positive: trips by bike moderately positive: car-sharing, leisure, shopping and private purposes	24.2 %

		moderately negative: household car and pt usage, accompanying, goods transport and work-related trips	
Multimodal users for all purposes (MUP)	highly mobile multimodal users	strongly positive: all trip purposes moderately positive: trips by car and by bike no negatively contributing variables	20.0 %

2.2. Random Forest model

Random Forest (RF) is a classification methodology that is based on a Classification and Regression Tree algorithm (CART) Breiman et al. (1984). The nodes of the generated tree represent binary decision rules which recursively split the data on the feature that maximises the information gain (IG) at each split. The IG is defined as follows:

$$IG(D_p, f) = I(D_p) - \sum_{j=1}^{m} \left(\frac{N_j}{N_p} I(D_j) \right)$$

Here, D_p and D_j are the data sets at the parent and the j-th child node, f is the variable to perform the split of the feature space. I is a function of the impurity (in our case the Gini impurity) whereas N_p and N_j are the number of samples at the parent and child node. The Gini impurity reflects the probability of misclassifying an observation. Formally, it can be described as:

$$GINI(\mathrm{D}) = 1 - \sum_{i=1}^{k} p_i^2$$

where D is the dataset containing samples of k classes and p_i is the proportion of the samples that belongs to class i for a particular node. Therefore, the Gini impurity is maximal if the classes are perfectly mixed.

The IG can be described as the difference of the impurity of the parent node and the sum of impurities at the child nodes. Thus, iteratively splitting the feature space based on the IG generates classes that are as pure as possible.

The RF-algorithm Breiman (2001) utilises an ensemble of classification trees, which means that instead of fitting one single tree to an associated classification problem, the algorithm fits an ensemble of so called weak learners (trees with reduced complexity) to bootstrap samples (random samples of the data set drawn with replacement) to achieve more robust and accurate results. In order to handle multicollinearity in the input data, the algorithm selects random features at each candidate split without replacement and therefore ensures the consideration of correlated strong predictors. Based on the outcomes of all weak learners, the class label is assigned by plurality voting (the class label which receives the most votes). Hyperparameter tuning of the RF-algorithm has been realized using a randomized search algorithm, a procedure in which random combinations of a predefined set of tuning parameters have been tested and evaluated to find the best-fitting combination. Evaluated hyperparameters are the maximum depth of the tree, the minimum number of samples required to split, the minimum number of samples required to be at a leaf node, the number of features to consider when looking for the best split and the number or trees in the forest. For this work 500 combinations of tuning parameters have been evaluated. Moreover, RF allows to find most relevant features (feature importance) for the respective classification task by computing the averaged impurity decrease (or IG increase) caused by each feature. To get an impression on how utilisation of different types of parameters effects the quality of the outcomes, three different parameter combinations were tested (see Table 2). First, the classification was performed with socio-demographic parameters only. Second, trip frequencies for specific modes were included additionally. Since we have mode-specific frequencies in both surveys of this study, this is the model we further apply for the data fusion. Third, accumulated trip frequencies of different purposes have been added to the model to show the potential

of imputing a dataset that includes trip frequency information, e.g. a tracking-based survey. For details on how trip frequencies were generated on the basis of the survey see Oostendorp et al. (2019).

In the next step, the model was used to assign any person of the national survey to a certain mobility type based on its reported socio-demographic parameters and trip frequencies with certain modes of transport.

Table 2: Groups of parameters used in the model.

Parameter group	**Description**
Socio-demographic attributes	Age, sex, persons in household, household type, number of cars in household, car-sharing membership, pt-ticket availability
Trip frequencies	Cumulated frequencies of Bike, pt, car (own), car (other), car-sharing trips
Trip purposes	Trip purposes divided by work, work-related, leisure, shopping, private, person transport, transport of goods

2.3. Evaluation

To perform the validation of the proposed method, the data set was split into 80 % training and 20 % testing data using stratified sampling. All predictions were evaluated by using a nested cross-validation technique in which the training data was split in three random folds that can be used to calculate an average F1-Score. F1-Score is defined as the harmonic mean of precision and recall where precision describes the proportion of positive classifications that are actually correct:

$$\textbf{Precision} = \frac{\textbf{True positives}}{\textbf{True positives} + \textbf{False positives}}$$

and recall is the proportion of actual positives in the dataset that are positively classified:

$$\textbf{Recall} = \frac{\textbf{True positives}}{\textbf{True positives} + \textbf{False negatives}}$$

This procedure has been performed for random combinations of possible tuning parameters to find the optimal hyperparameter configuration (see section 2.2). Finally, the predictor with the best fit was applied to the testing sample to evaluate the performance on unseen parts of the data. The success of the imputation from the initial data set to the other is assessed by comparing the distribution of the share of the individual user groups in Berlin and the overall sample of the MiD in comparison to the initial data set.

3. Results

This section demonstrates the result of classifying single persons of travel surveys into pre-defined mobility types using the methods described in the previous chapter. It is divided into two parts: The first part shows the results of the model prediction based on a local survey conducted in Berlin (see 2.1.1). It illustrates how the use of different types of socio-demographic and trip-specific parameters effects the quality of the classification considering three different groups of parameter combinations (see Table 2). The second part shows the outcomes of the transfer of the pre-defined user typology to the subset of people of the national travel survey MiD living in Berlin. This allows a comparison of percentages in the two surveys.

3.1. Berlin survey

The Berlin survey is the basis for the training and the evaluation of the model. Figure 3 shows the result of the classification using different types of input parameters. Corresponding to the identified mobility types in the Berlin survey (see 2.1.2), the target classes are named public transport users (PTU), bicycle centred users (BCU), lower activity car users (LACU) and multimodal users for all purposes (MUP). Using only socio-demographic parameters for the prediction of mobility types leads to a poor average cross-validated F1-score of 0.46. By adding trip frequencies to the model, this value increases to 0.82. When including trip accumulated trip frequencies and cumulated trip purposes, the F1-score increases to 0.95.

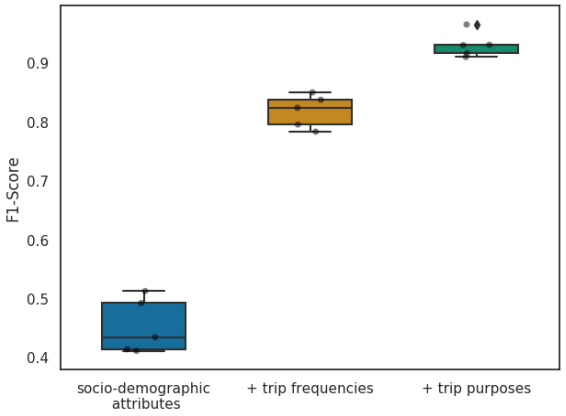

Figure 3: Effects on the quality of the classification of mobility types using different combinations of input variables (see Table 1). The black dots illustrate the result of the respective cross-validation step. Mean f1-scores are 0.46 (socio-demographic attributes), 0.82 (socio-demographic attributes + trip frequencies) and 0.93 (socio-demographic attributes + trip frequencies + trip purposes).

The application of exclusively socio-demographic variables for classification leads to a poor allocation for all classes. When looking at the results of the testing data set (part of the data that has not been used for the training of the model), the highest F1-scores show PTU and LACU with 0.46 and 0.42. MUP and BCU reach 0.33 and 0.39 (see Table 3**Fehler! Verweisquelle konnte nicht gefunden werden.**). Most important variables are here the availability of public transport ticket (0.3 averaged IG increase), followed by number of cars in the household (0.21 averaged IG increase), age (0.18 averaged IG increase), and the number of persons in the household (0.14 averaged IG increase). Car-sharing membership, sex and household type play a less important role in the model (see Figure 4a).

Table 3: Results of the classification of mobility types (testing data set)

	socio-demographic attributes			*+ trip frequencies*			*+ trip purposes*		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
LACU	0.47	0.38	0.42	0.88	0.89	0.88	0.98	0.95	0.96
PTU	0.38	0.59	0.46	0.85	0.94	0.89	1.00	0.98	0.99
BCU	0.45	0.34	0.39	0.77	0.80	0.78	0.92	0.97	0.94
MUP	0.31	0.34	0.33	0.56	0.46	0.51	0.88	0.88	0.88

When adding additional accumulated trip frequencies, the F1-scores considerably increase for all classes. Again, the best results show LACU and PTU with 0.88 and 0.89 followed by BCU (0.78) and MUP (0.51) (see Table 3). Here, trip frequencies of public transport, bike and car trips (0.29, 2.9 and 0.11 averaged IG increase) are more

important than the socio-demographic attributes of the respective person (see Figure 4b). Additionally, the accumulated frequency of all trips and the accumulated car-sharing trips are less important influences.

Best results show the outcomes of the classification incorporating accumulated trip frequencies and accumulated trip purposes. In this case, the classes PTU, LACU and BCU are nearly perfectly assigned (f1-scores of 0.99, 0.96, 0.94). Even the least accurate class MUP shows an acceptable score of 0.88 (see figure in Table 3). Most influencing factors are again the number of public transport and bike trips (0.23 and 1.9 averaged IG increase) followed by accumulated trips for goods transport and accompanying tips, private trips and car trips. Socio-demographic variables are again less important than trip frequencies (see Figure 4c).

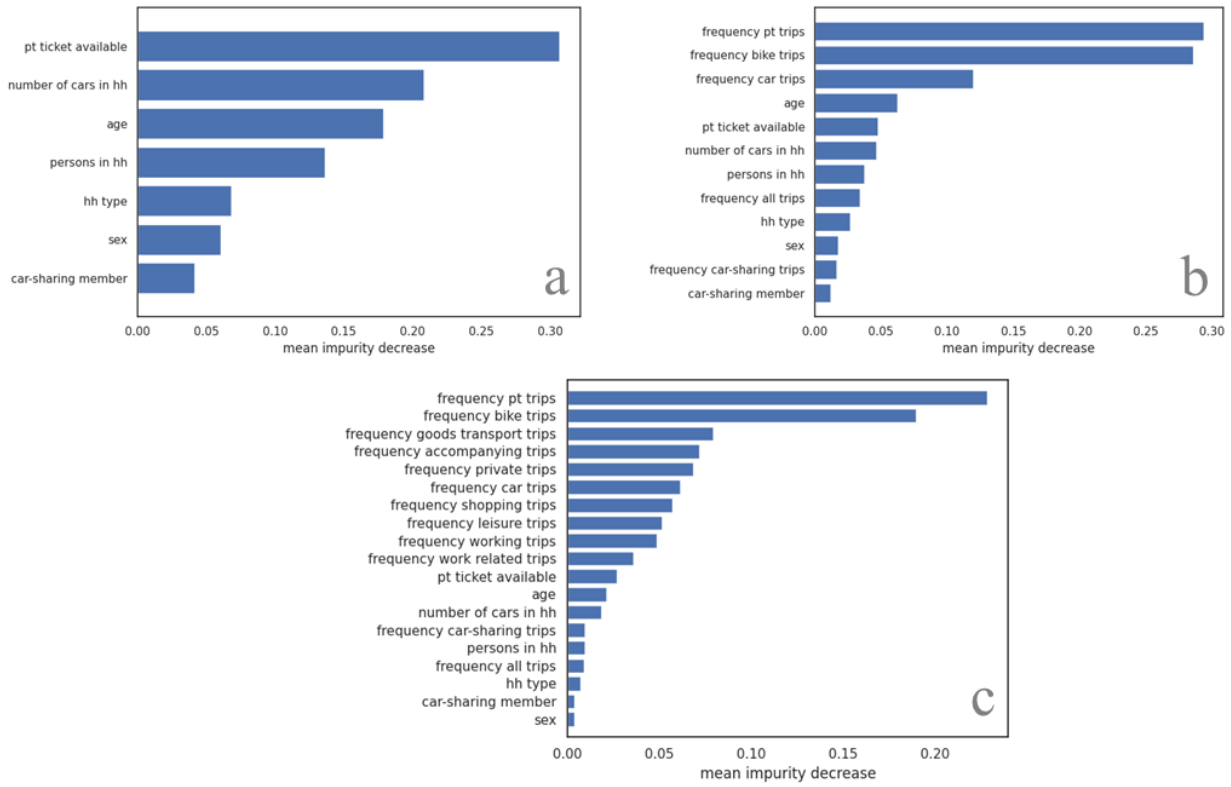

Figure 4: Feature importance of the three models: socio-demographic attributes only (a), socio-demographic attributes + trip frequencies (b), and socio-demographic attributes + trip frequencies + trip purposes (c).

3.2. Transfer to (Berlin part of) the national household travel survey MiD

Figure 5 summarizes the results of the transfer of pre-defined mobility types (see section 2.1.2) from a local survey of Berlin into the national household travel survey of Germany MiD. Based on a RF-classification approach using socio-demographic variables and accumulated frequencies of trips conducted by bicycle, public transport, car and car-sharing, each person in the MiD was assigned to one of the defined mobility types. The figure illustrates the share of persons belonging to the four mobility types LACU, PTU, BCU, MUP according to three different data sources: First, the blue bars show the actual share of each mobility type according to the Berlin survey. Second, the orange bars show the results of the prediction for the testing data , the part of the Berlin survey that has not been used for the training of the model. Finally, the green bars illustrate the percentages of each mobility type in the Berlin part of the MiD resulting from the application of the RF-model.

As shown, shares for the class LACU differ only slightly in the three data sets (34.7 %, 35.6 % and 35.2 %), indicating fairly good results for the application of the transfer model. Noticeably, compared to the original data, the shares are slightly overestimated in the testing data, but underestimated in the MiD. In contrast, class membership

both for PTU and BCU are higher in both prediction data sets than in the source data. For the PTU-class, the original share of 20.2 % is contrasted by 21.9 % in the testing data and 24.9 % in the MiD. BCU-shares, originally amounting to 24.4 % in the local data set, rise to 25.9 % in the testing data and further to 27.6 % in the MiD case. The larges deviance between actual and predicted shares of class membership are observable for MUP. While 20.2 % of the respondents in the source data exhibit this behaviour, shares predicted for the Berlin testing data (16.6 %) and the Berlin-sample of the MiD survey (12.8 %) are significantly lower.

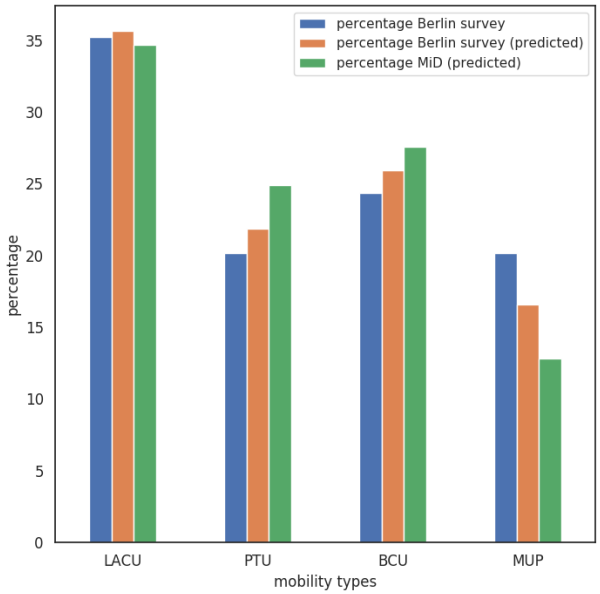

Figure 5: Comparison of class percentages of mobility types in the Berlin survey and the Berlin part of the national survey MiD.

4. Discussion

The results show that transfer of information from a highly specialized survey to a more general, bigger national household travel survey is possible using a random forest classification methodology. The results of the validation show that exclusively using socio-demographic attributes for the transfer lead to poor results and can therefore not be recommended. The best results in this case can be detected in the classes PTU and LACU, classes that are characterized by one-sided behaviour patterns and the ownership of mobility resources (pt-ticket, car) that support regular usage of certain modes.

A particular challenge is the correct classification of MUP-membership. While the shares for all other groups are approximately met or slightly overestimated, class-membership for the MUP-type is considerably underestimated. In the original dataset (local survey) MUP is characterized by many positive contributing variables (e.g. all trip purposes are strongly positive), no negative contributing variables, that means many factors influence whether a person is attributed to this group (see Table 1). As a consequence, the group is very heterogeneous and therefore more difficult to be predicted or described by trip frequencies of certain modes or purposes. For PTU and BCU there is only one strongly contributing variable and, additionally, this is the trip frequency of one transport mode (namely, pt and bike, respectively) (see Table 1).

In general, the model quality significantly rises when information on actual behaviour is included. Therefore, there is a clear improvement by involving accumulated trip frequencies of different transport modes. By additionally adding

trip purposes, the results increase to an F1-score of 0.95, almost twice as high as the initial socio-demographics-only model.

One difficulty of the approach lies in the validation of the output data. The more variables are used for generating a transfer model, the less variables are available for verification. Saving one possible input variable for verification lead to the dilemma that either the variable was not very distinct in the different classes like possession of driver license or a key input variable like possession of public transport ticket. Our work uses the frequencies of each class to verify the data. The qualitative ranking of the classes is fit very well. Interestingly, the frequencies are fitted best for the classes BCU and LACU, where ownership of the bike or car is a key factor to be member of this class. However, bike ownership is very common even in pt- and car-centred mobility groups. Indeed, analysis of the MID data set shows that about 76 % of German households own at least one bike (Nobis, Kuhnimhof, 2018) that can be a flexible transport option, especially as fall-back and for short distances.

As shown by Shah et al. (2014) and Stekhoven and Bühlmann (2012) for other disciplines, we can conclude that random forests are highly suitable for data fusion tasks in travel surveys. because of its non-parametric nature and ability to automatically detect relevant features it can be especially recommended for data fusion tasks that are based on a high number of relevant and possibly highly correlated variables. The main drawback of the methodology is that the effects of individual variables on the transfer can only be derived limitedly and a knowledge-based implementation is not possible.

5.Conclusion

This paper illustrates that random forest classification models are suitable for transferring information from highly specialized surveys to more general travel surveys, often collected on national level. In the case presented, four mobility types developed using a small, local data set collected in Berlin were transferred to the German national household travel survey „Mobility in Germany“ (MiD) on the basis of mobility behaviour parameters present in both surveys. The RF methodology is a non-parametric method for classification which does not make assumptions about the distribution of the variables, can deal with any kind of data types (categorical, continuous and mixed-types), is able to solve classification problems with strongly non-linear relationships, is easy and fast to implement and be applied without domain knowledge. It therefore represents a recommendable alternative to exiting methods. However, the added value of the merged content and the quality of the transfer procedure depend on the depth and overlap of the information used for the classification modelling. As shown, relying only on socio-demographic information typically available in mobility surveys for linking source and target survey might not be sufficient. This calls for the identification of research-question-specific variables in the target data set prone to serve as linkage variables ideally already when designing the specialized survey. Their inclusion in the source survey is likely to positively influence the quality of the random forest classification model and hence the transferability of the detailed results. In our case, the transferability of the mobility types is strongly relying on aggregated frequencies of inter- and unimodal usage characteristics. Information on recurring travel mode choice is often gathered using tracking-based surveys or, typically on a more aggregate level, big data sources. Therefore, the transferability analysis presented also shows possibilities how traditional travel surveys can be combined with new data sets based on tracking or big data.

References

Bahamonde-Birke, F.J., Hanappi, T., 2016. The potential of electromobility in Austria: Evidence from hybrid choice models under the presence of unreported information. Transportation Research Part A: Policy and Practice 83, 30-41.

Bayart, C., Bonnel, P., Morency, C., 2009. Survey mode integration and data fusion: methods and challenges. In: Bonnel, P., Lee-Gosselin, M., Zmud, J., Madre, J.-L. (Eds.), Transport survey methods: Keeping up with a changing world. Emerald Press, pp. 587-611.

Breiman, L., 2001. Random forests. Machine learning 45, 5-32.

Breiman, L., Friedman, J., Olshen, R., Stone, C., 1984. Classification and regression trees. Wadsworth Int. Group 37, 237-251.

Chlond, B., 2013. Multimodalität und intermodalität. In: Beckmann, K.J., Klein-Hittpaß, A. (Eds.), Nicht weniger unterwegs, sondern intelligenter? Neue Mobilitätskonzepte. Edition Difu, Berlin, pp. 271-294.

Eggs, J., Follmer, R., Gruschwitz, D., Nobis, C., Bäumer, M., Pfeiffer, M., 2018. Mobilität in Deutschland− MiD: Methodenbericht. Studie von infas, DLR, IVT und infas 360 im Auftrag des Bundesministeriums für Verkehr und digitale Infrastruktur (FE-Nr. 70.904/15). Bonn, Berlin.

Goulias, K.G., 2000. Surveys using multiple approaches. Transport Surveys: Raising the Standard, Grainau, Germany.

Haustein, S., Hunecke, M., 2013. Identifying target groups for environmentally sustainable transport: assessment of different segmentation approaches. Current Opinion in Environmental Sustainability 5, 197-204.

Heldt, B., Donoso, P., Bahamonde-Birke, F., Heinrichs, D., 2018. Estimating bid-auction models of residential location using census data with imputed household income. Journal of Transport and Land Use 11, 1101-1123.

Hildebrand, E.D., 2003. Dimensions in elderly travel behaviour: A simplified activity-based model using lifestyle clusters. Transportation 30, 285-306.

Hunecke, M., Haustein, S., Böhler, S., Grischkat, S., 2010. Attitude-based target groups to reduce the ecological impact of daily mobility behavior. Environment and behavior 42, 3-43.

Jensen, M., 1999. Passion and heart in transport—a sociological analysis on transport behaviour. Transport Policy 6, 19-33.

Jones, W.B., Cassady, C.R., Bowden Jr, R.O., 2000. Developing a standard definition of intermodal transportation. Transp. LJ 27, 345.

Krizek, K.J., Waddell, P., 2002. Analysis of lifestyle choices: Neighborhood type, travel patterns, and activity participation. Transportation research record 1807, 119-128.

Nobis, C., Kuhnimhof, T., 2018. Mobilität in Deutschland− MiD: Ergebnisbericht. Studie von infas, DLR, IVT und infas 360 im Auftrag des Bundesministers für Verkehr und digitale Infrastruktur (FE-Nr. 70.904/15). Bonn, Berlin.

Oostendorp, R., Gebhardt, L., 2018. Combining means of transport as a users' strategy to optimize traveling in an urban context: empirical results on intermodal travel behavior from a survey in Berlin. Journal of Transport Geography 71, 72-83.

Oostendorp, R., Nieland, S., Gebhardt, L., 2019. Developing a user typology considering unimodal and intermodal mobility behavior: a cluster analysis approach using survey data. European Transport Research Review 11, 33.

Outwater, M.L., Modugula, V., Castleberry, S., Bhatia, P., 2004. Market segmentation approach to mode choice and ferry ridership forecasting. Transportation Research Record 1872, 71-79.

Prillwitz, J., Barr, S., 2011. Moving Towards Sustainibility? Mobility styles, attitudes and individual travel behaviour. Journal of Transport Geography 19, 1590-1600.

Shah, A.D., Bartlett, J.W., Carpenter, J., Nicholas, O., Hemingway, H., 2014. Comparison of Random Forest and Parametric Imputation Models for Imputing Missing Data Using MICE: A CALIBER Study. American Journal of Epidemiology 179, 764-774.

Stekhoven, D.J., Bühlmann, P., 2012. MissForest—non-parametric missing value imputation for mixed-type data. Bioinformatics 28, 112-118.

Van Buuren, S., Oudshoorn, K., 1999. Flexible multivariate imputation by MICE. Leiden: TNO.

Waljee, A.K., Mukherjee, A., Singal, A.G., Zhang, Y., Warren, J., Balis, U., Marrero, J., Zhu, J., Higgins, P.D., 2013. Comparison of imputation methods for missing laboratory data in medicine. BMJ open 3.