# IDENTIFYING TRIP PURPOSES ON TRIP LEVEL FOR VEHICLE SENSOR DATA

**Ulrich Niklas (corresponding author)**
BMW AG
Petuelring 130, 80788 Munich,
Email: ulrich.niklas@bmwgroup.com

**Miriam Magdolen**
Institute for Transport Studies, Karlsruhe Institute of Technology (KIT)
Kaiserstrasse 12, 76131 Karlsruhe, Germany
Email: miriam.magdolen@kit.edu

**Sascha von Behren**
Institute for Transport Studies, Karlsruhe Institute of Technology (KIT)
Kaiserstrasse 12, 76131 Karlsruhe, Germany
Email: sascha.vonbehren@kit.edu

**Peter Vortisch**
Institute for Transport Studies, Karlsruhe Institute of Technology (KIT)
Kaiserstrasse 12, 76131 Karlsruhe, Germany
Email: peter.vortisch@kit.edu

Word Count: 6,858 words + 2 tables (250 words per table) = 7,358 words

*Submitted August 1, 2020*

## ABSTRACT

The understanding of car usage patterns and the reason for a trip is important for policymakers to derive measures to influence car usage as well as for manufacturers and service providers to create target-oriented products and offers. There are different types of data to describe car usage. Survey data provide various information that explain behavior of individuals for a short time period. In contrast, sensor data from cars contain detailed usage data over a longer time period, but do not allow conclusions to be drawn about the purpose of the trip. In existing literature is a lack of research on how to combine the best of both data types without explicit validation by participants. The aim of this study is to identify and analyze the purposes of trips in sensor data by using a car use model that is based on survey data from a national household travel survey. The characterization of trips with different purposes is used to train a model. This is applied to sensor data of about 51,000 cars from nine European countries with 7,489,686 trips in the course of half a year of a German premium Original Equipment Manufacturer. The results show that the chosen approach is useful for the identification of trip purposes in sensor data. All in all, 73% of the trip purposes in the model could be correctly predicted at trip level. In an in-depth analysis, we compare car usage across the nine countries considered and evaluate the trips differentiated by fuel type.

**Keywords:** Europe, trip purpose, sensor data, survey data, car usage

## INTRODUCTION

The recording, investigation and the resulting understanding of car usage is a challenge for stakeholders. This is especially true for policymakers to derive measures to influence car usage, but also for manufacturers and service providers to create target-oriented products and offers. Cars from different car segments are used differently by the users, which is why the cars must meet different needs. For example, some cars are used primarily for long-distance trips, while others are used only for short distances in everyday life or for daily commuting. The investigation of car usage therefore allows to develop political measures that influence a certain aspect of the usage behavior. To make distinctions between the car use behaviors, the investigation of car travel data is necessary. However, such data sources are rare.

In many cases, car use behavior is investigated with the help of survey data from National Household Travel Surveys (NHTS). Thereby it is possible to consider the reported trips by car of the respondents in the survey. In addition to the attributes of the individual trip, other important information is available. This includes the socio-demographic characteristics as well as the activities of the participant, i.e. the trip purposes. The information is thus available at individual level and not at car level. Only little information is given about which car is used. As there are no specific trip diaries for cars, it is not possible to identify the usage profile of a specific car in the household. Especially, when several people share more than one car in the household, each with a different usage pattern. Furthermore, NHTS from different countries have different survey concepts, e.g. survey period (cross-sectional or longitudinal) or survey instruments and are therefore not directly comparable with each other.

Next to travel survey data, sensor data of cars allow to describe car usage. This type of data provides comprehensive information on the use in time and space and allows the differentiation of usage types. An advantage of sensor data is that the collection of the information is not leading to a respondent burden. Data collection faces only technical limitations, e.g. poor reception or data transmission. The data are collected automatically and there are no constraints or limitations for the users of the cars. However, there is the drawback that important information on travel behavior cannot be recorded with sensor data: The purpose of trips. With the information on time and space, it is not possible to directly determine the destinations, i.e. also the activity at the location and thus the trip purposes. In general, information from the spatial structure, for example the Points of Interests (POI) can be used to determine the activity at the destination. However, the spatial information is often not precise or distinct, especially in city centers. Furthermore, some trip purposes can also be determined via the regularity of usage (e.g. work, home), but are mainly based on assumptions.

Both, survey and sensor data have specific advantages and contain information that is missing in the other data source. It is still unclear how a link between the two data sources can be achieved without making rough assumptions or losing information. How can sensor data be enriched with the information on the trip purpose? How can such extended data be used and what information and insights can be drawn from such extended data?

In this paper we present an approach of using travel survey data to estimate trip purposes recorded in sensor data. For this, we use a data set from a car use model, which is based on longitudinal survey data from the German Mobility Panel (MOP). By analyzing the characteristics of trips with different purposes, such as the time of the day, insights are gained and the usage of cars differentiated by trip purpose is described. The model also provides information about the standing time of the vehicle, which is particularly important for the identification of places of residence. The described data are used to train a model which is applied to sensor data to identify the purposes of the trips. For this analysis, a large data set with sensor data of about 51,000 cars of a German premium Original Equipment Manufacturer from nine European countries is used.

The paper is structured as follows: First, we present a brief literature review on the methods for identifying trip purposes for sensor data. This is followed by a description of the data used for our analysis. Third, we present the approach used to identify and complement the trip purpose information to the available data and to create an extended data set on car use. The results are presented and analyses of differences in car usage between countries and between different fuel types are discussed. Finally, we discuss the limits of our approach, draw a conclusion and refer to further work.

## LITERATURE REVIEW

Several studies deal with the identification of trip purposes or respectively of destinations and the activities at the destination. This issue is particularly addressed in the literature on the analysis of mobile phone data. Alexander et al. (*1*) use a rule-based approach to identify the location of home and work of about 2 million mobile phone users. The location of home is defined as the place with the most stays on weekends and on weekdays between 7 p.m. and 8 a.m. Calabrese et al. (*2*) also use mobile phone data to extract information on the travel behavior of individuals. Again, a rule-based approach determines the location of home. It is considered in which defined cells users are connected to the network between 6 p.m. and 8 a.m. The cell in which the user is connected most nights is selected as home location. Isaacman et al. (*3*) give another example for identifying the important locations of people by analyzing mobile phone data. Activities at home and work locations explain a relevant part of the people's time use. In this study, about 170,000 mobile phones are analyzed. Since a small number of participants reported the actual place of home and work, this information served as the basis for a logistic regression. This was used to determine the place of home and work for the remaining data. As a result, the home location was chosen where most hours were spent at weekends and on workdays in the period between 7 p.m. and 7 a.m. The work location is described by the location with most stays in the period from 1 p.m. to 5 p.m.

The studies mentioned above show that rule-based definitions can be used to describe the location of home and work. Trips that end at such places can therefore be identified as trips to work and trips back home. For the identification of other important locations and thus also other trip purposes, such as leisure or shopping, different approaches are necessary. Some studies use spatial data including information on land use and POIs. Such POIs are for example restaurants, hotels, or supermarkets. Phithakkitnukoon et al. (*4*) divide their study area into smaller grids and define the type of activities that are likely to be performed in this grid by using the information of POIs. For example, a grid with a city park has the activity “recreation”. From this information for all visited grids by a mobile phone user, daily activity patterns are created.

Next to mobile phone data, Global Positioning System (GPS) data is common in the field of travel behavior research. Gong et al. (*5*) give a comprehensive overview on literature on existing methodologies to identify trip purposes in GPS data. Bohte and Maat (*6*) perform an approach that combines GPS data, Geographic Information System (GIS) data and a web-based validation application to identify trip purposes. The purposes of the trips are derived by drawing a 50-meter radius around the destination. If a POI is within this radius, it is assumed that this POI has been visited. If there are several POIs within the radius, the nearest POI is used. This demonstrates a weakness in the approach. As soon as several POIs are available within a small radius, no clear assignment is possible. Seo at al. (*7*) use an approach in which on the one hand the purpose of the trip is estimated and on the other hand the participants are asked about the actual trip purpose. By knowing the actual trip purposes, a learning process can be implemented. A further example for the combination of GPS data and the validation of the trip characteristics by the participant is given by Xiao et al. (*8*). However, in these studies the trip purposes are collected and available in the data set. Furthermore, the sample for these studies are small, since the data collection of the GPS information and the specification of further information such as the trip purposes is demanding and expensive.

When reviewing the literature and discussing the methodologies, a distinction must be made between trips made by people, e.g. those reported in a trip diary, and trips of vehicles, i.e. vehicle usage data. Travel surveys, e.g. the NHTS in the USA (*9*) or the MOP in Germany (*10*), have the advantage of recording not only travel behavior data but also characteristics of the participants. These include characteristics that are considered to influence travel behavior, such as age or income. Furthermore, the activities and with this also the purposes of trips, are reported by the participants themselves. There are studies that use the data from such surveys in order to transfer them to vehicle usage data and thus identify the purposes of the trips. Leerkamp (*11*) use survey data from the German NHTS "Mobilität in Deutschland" (*12*) to determine the characteristics of trips with different purposes. Such characteristics are trip distance, starting time of the trip and trip duration. This information is then used in the identification of trip purposes of sensor data. However, a shortcoming of using data from a cross-sectional NHTS is, that the period of data collection is short and some trip purposes, such as long-distance travel are often not comprehensively captured. Further, survey data describes the travel behavior of individuals and a direct transfer of purposes of trips by car as a driver or as a passenger to car usage data is not possible. Survey data do not provide information on how cars are used and for which purposes they are used across different countries, e.g. to investigate whether certain fuel types are increasingly used for specific trip purposes. If there are several cars in a household, it is not clear from the survey data which cars are used for which purposes. At the same time, several people may use the same car, which makes the car usage pattern more complex than the survey data reveal.

For the reasons mentioned above, it is preferable to use car usage data for the identification of trip purposes in sensor data. However, car usage data is rare. The Car Usage Model Integrating Long-Distance (CUMILE) (*13; 14*) describes the car use of vehicles for a period of one year, including detailed trip characteristics. In a previous study, Niklas et al. (*15*) used sensor data to compare car usage profiles of electric vehicles with those from conventional vehicles. Car usage profiles that resulted from a cluster analysis based on the data from CUMILE were used to describe the conventional vehicles. With a probabilistic approach, the electric vehicles in the study were allocated to the car usage profiles. In a first attempt, the study of Niklas et al. (*15*) emphasizes that sensor data and the data from the CUMILE model are comparable. In this present paper, we also use these two types of data sources, which we describe in more detail in the following section.

## SENSOR DATA

The sensor data available for this study was collected from a German premium Original Equipment Manufacturer between 2019-10-17 and 2020-05-26. 50,858 vehicles made 7,489,686 trips in this period, which represents a large data set for the following analyses. The sensor data includes vehicles from nine European countries, namely Austria, Belgium, Denmark, France, Germany, Italy, Spain, Netherlands, and the UK. For each trip, information is available on an anonymized vehicle identification number, trip ID and, in addition, the date and time are recorded according to a timestamp. The country in which the dealer is located is used to assign the cars to countries. Due to data protection reasons it is only possible to conclude the country. Further variables are prepared from the given information. For the following analyses, the start and end of the trip and the resulting duration of the trip are relevant. Furthermore, the difference between the end of a trip and the beginning of the following trip is used to calculate the duration of the standing time of the vehicle. This information plays a decisive role for the identification of the trip purposes, as will be shown in the following.

## SURVEY DATA

As ground truth for the identification of the trip purposes a unique data set is available from the car usage model CUMILE (*13; 14; 16*). This model is based on the survey data of the German Mobility Panel (MOP) and describes the car usage of households over a period of one year. The MOP is a household travel survey and consists of two types of surveys and therefore offers two types of data sets. The survey on 'Everyday Mobility' collects the travel behavior of the participants over a period of one week by means of a trip diary. In the survey on 'Fuel Consumption and Odometer Reading' the members of the household fill in the information for the vehicles each time they refuel over a period of eight weeks. Ecke et al. (*10*) gives a detailed description of the two survey parts of the MOP. For the modelling of car usage in CUMILE, the travel behavior of all household members is considered. Trips by car recorded in participant's trip diaries is transferred to the cars in the household. This means that the purposes of trips by car are also transferred from the household members to the car, which results in a trip diary of the car. In addition, this makes it possible to better describe the standing times of cars, as the use of several people is taken into account. As an additional data source CUMILE includes the long-distance survey INVERMO. In this survey, participants reported retrospectively their last three long-distance trips (*17*). Eisenmann (*13*) modelled in CUMILE the car usage over one year by combining the information of all mentioned data sources. In total, 6,309 cars with 4,559,288 trips in Germany are included in the model and used for this study.

As the CUMILE data include car usage over a long period of time, we have a comparable data set of sensor and survey data to approximate the trip purposes in the sensor data. In addition, the combination of the CUMILE dataset and sensor data has already been implemented in previous studies (*15; 18*). In the following section, we will explain our methodological approach in more detail.

## ANALYZING CUMILE DATA

For the approximation of the trip purposes in the sensor data, we analyze the car use data from CUMILE. The trips in CUMILE are used to train a model that estimates the purpose of the trips in the sensor data. CUMILE includes 11 trip purposes. Since there is the need of a reliable database for the following estimation of the trip purposes, we decided to exclude the "outliers", i.e. trip purposes with a relative low number of trips, from the following analyses. This is the case for to the trip purposes 'education', 'other', 'second home' and 'loop trip', which each account for only 1.5% (70,000 trips) or less. The total number of the remaining trips is 4,455,750. The distribution of the trip purposes is shown in Figure 1. At this point we underline again that these are exclusively trips of cars.

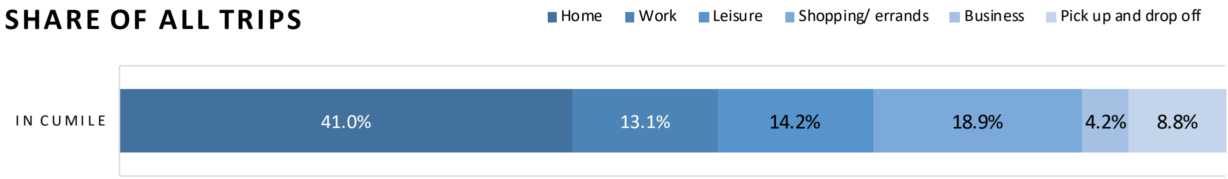

**Figure 1 Distribution of trip purposes in CUMILE**

Since the aim is to apply the model to the sensor data, we make sure that only input variables are used which can be generated from the sensor data. To determine the input variables for estimating the trip purposes at the trip level, we first visually examine the characteristics of the trips differentiated by purpose with different characteristics. Figure 2 shows the distribution of the standing time up to 1110 minutes differentiated by trip purpose of the trip before the standing time. This important information about the car is available in CUMILE. In regular travel behavior surveys, such information would not be provided. Differences between the trip purposes become clear. The longest standing times occur after trips with the purpose ‘home’. From the distribution it can be concluded that the longer the standing time, the more likely the previous trip has the purpose ‘home’. In addition, relatively long standing times after trips with the purpose ‘work’ are also evident. If the standing time is approximately about 240 to 630 minutes, the trip is likely to have the purpose ‘work’.

In the case of very short standing times, the trip purposes ‘shopping/ errand’ and ‘pick up and drop off'' are dominant. It is interesting for the purpose ‘shopping’ that the share increases again with a standing time of about 390 minutes. This effect is even stronger if only weekend trips are considered. Day trips with the purpose of shopping in a shopping center or in a larger city may be such events that explain the increase. Regarding the purpose ‘business’, the figure indicates that such trips take place with a standing time of at least about 30 minutes.

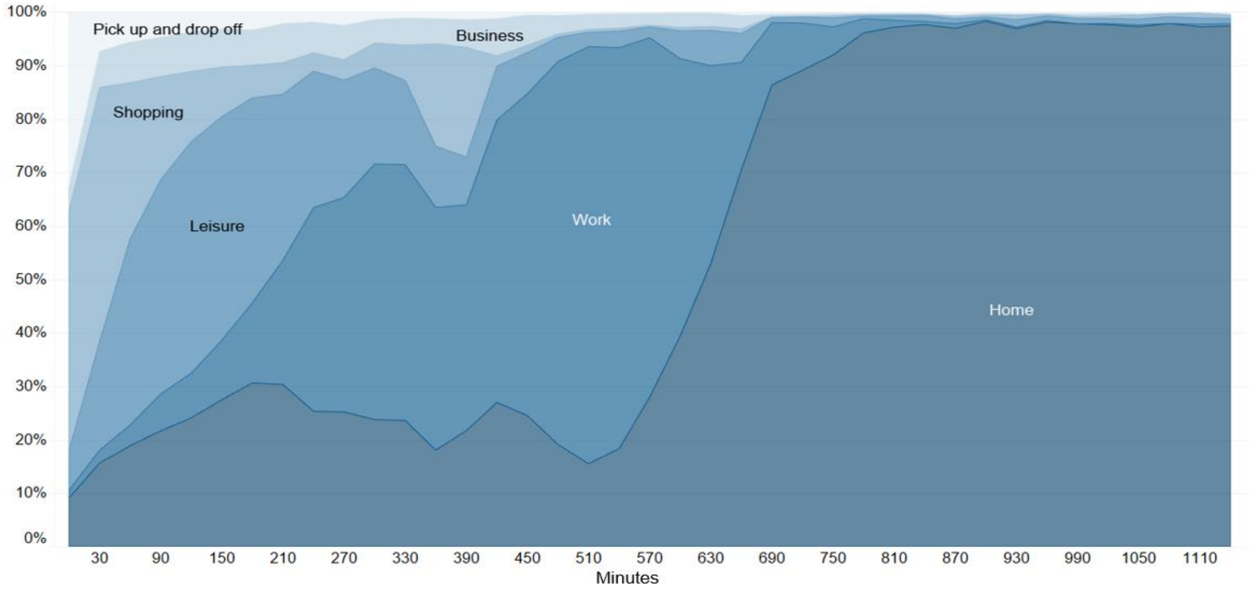

**Figure 2 Distribution of standing time differentiated by trip purpose**

Another characteristic of the trips in CUMILE, which is visualized in Figure 3, is the start time of the trip. In the early morning, the purposes of the trips are highly likely to have the purpose 'home'. Between 4 a.m. and 7 a.m. there are mainly trips to work in the data. However, the later the start time is on a day, the more likely it is that the trips have the purpose 'home'. Trips with the purpose 'shopping/ errands' take place mainly in the morning (8 a.m. to 11 a.m.). The purpose 'business' shows an outlier at 3 a.m. However, this should be treated with caution, as there are generally few trips (1,171 trips) at this time of the day. The share of trips with the trip purpose 'pick up and drop off' is relatively stable over time.

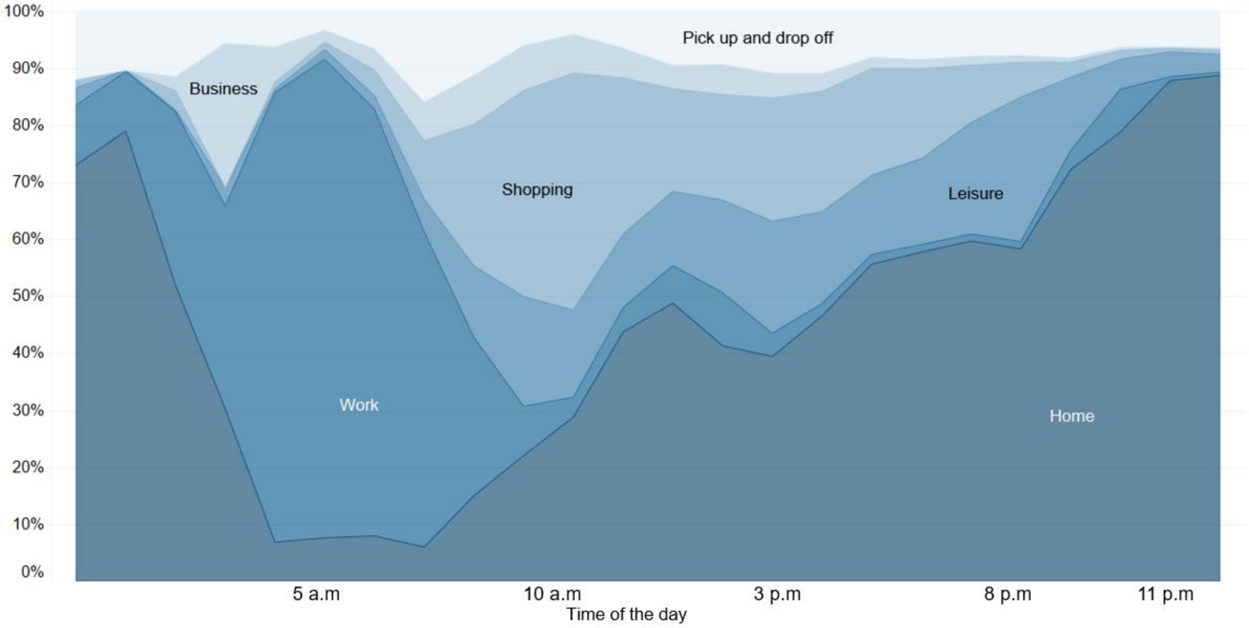

**Figure 3 Distribution of starting time differentiated by trip purpose**

The two preceding descriptive analyses were useful for a first evaluation and for the derivation of characteristics that may be relevant for the identification of trip purposes. In addition, the distances differentiated by trip purpose are included in the following analyses. Figure 4 shows the distribution of the trip distances. The figure indicates that the length of trips has no influence on the trip purpose and no clear findings can be demonstrated with this visualization. For this reason, the distances were not considered in the modelling.

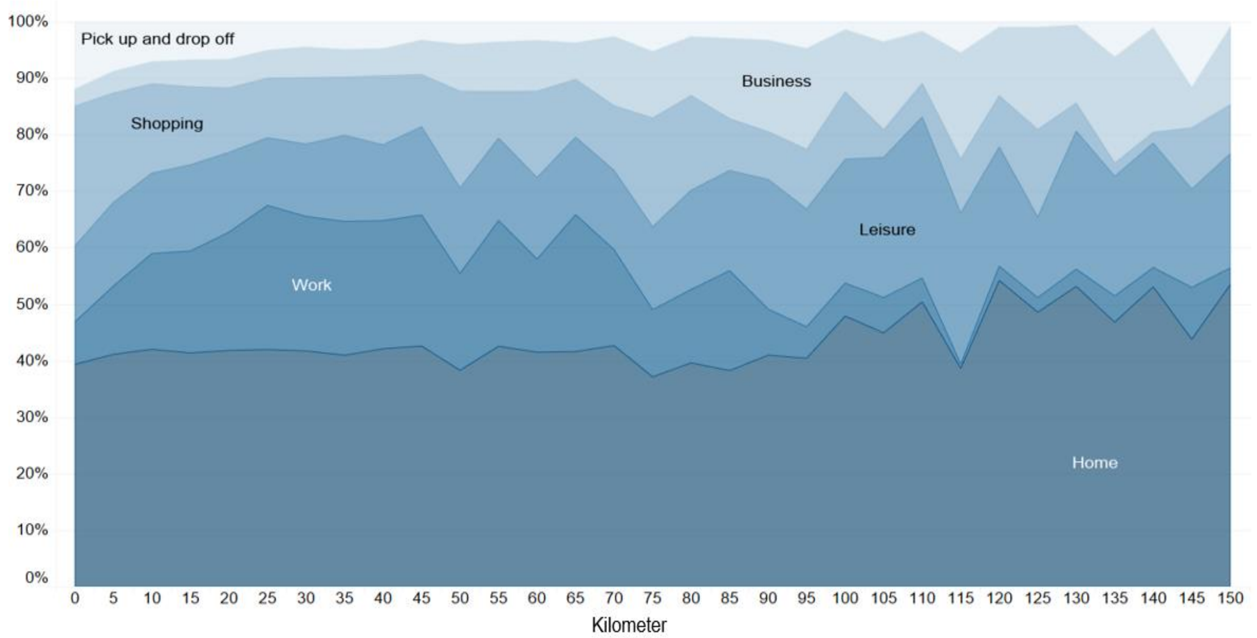

**Figure 4 Distribution of distances in kilometer differentiated by trip purpose**

## METHODOLOGY

In this section, we describe the applied approach to approximate trip purposes in the sensor data. The findings and results from the evaluation of CUMILE will be used to perform a supervised learning technique, called Softmax Regression (SR). First, we explain the used input variables. Next, we describe the SR approach. Finally, we validate the trained model.

### Input variables

The input variables consist of information on the standing time and start time. To enable the model to generate more precise decision functions, the continuous variables are divided into intervals to create binary variables. Figure 2 and Figure 3 serve as indications for setting the interval limits. The standing time is grouped into seven intervals (0-30, 30-60, 60-120, 120-240, 240-360, 360-480 minutes and above 480 minutes). The start time of the trips is also divided into seven groups, whereby the group in the early morning captures more hours than the others during the day (midnight-6 a.m., 6 a.m.-9 a.m., 9 a.m.-noon, noon-3 p.m., 3 p.m.-6 p.m., 6 p.m.-9 p.m. and 9 p.m.-midnight). In addition, the information of a trip is not only processed at the time *t* but also at *t-1* and *t+1*. This is relevant for the classification since several alternatives were tested in the process of variable selection. Thus, the starting time of a trip at the times *t* and *t-1*, as well as the standing time after a trip at the times *t*, *t-1* and *t+1* were considered. In addition, the information whether the trip took place on a working day and whether the car spent the night after the trip was taken into account. All in all, we generated 37 variables to characterize the purpose of a trip. After the preprocessing steps (deletion of trips with missing values or zeros), a total of 4,398,678 trips were generated with the above-mentioned information.

### Softmax Regression model

We use the purpose of each trip $i$ as label and the generated variables regarding standing time and start time as features from CUMILE $C = \{(\boldsymbol{x_i}, y_i)\}_{i=1}^{4,398,678}$. $\boldsymbol{x_i}^{(l)}$ is a matrix representing the 37 features for trip $i$. $y_i^{(l)}$is a vector representing the trip purpose for trip $i$. The aim of the model is to generate a softmax (decision) function $\boldsymbol{f}(\boldsymbol{x}_i^{(l)}) = y_i^{(l)}$ by adjusting the parameters $\theta$, which describes the relation between $\boldsymbol{x_i}^{(\boldsymbol{l})}$ and $y_i^{(l)}$ in the most accurate way. The lbfgs (Limited-memory Broyden–Fletcher–Goldfarb–Shanno) algorithm was used to adjust the parameters $\theta$ and optimize the softmax function (*19*). A softmax regression, where multiclassification problems can be solved, is a generalization of the logistic regression (*20*). To measure the accuracy of the model, the 4.398.678 trip are randomly split into a training $C^{train} = \{(\boldsymbol{x}_{\boldsymbol{i}}^{\boldsymbol{train}}, y_i^{train})\}$ (67%) and test data set $C^{test} = \{(\boldsymbol{x}_{\boldsymbol{i}}^{\boldsymbol{test}}, y_i^{test})\}$ (33%). Test data is only used for validation. With the training data the parameters $\theta$ are optimized and the decision function is generated. To ensure generalization, the decision function for estimating trip purposes is applied to the test data set. The predicted trip purposes $\hat{y}_i^{test}$ of a trip are now compared with actual purpose $y_i^{\boldsymbol{test}}$ of a trip.

### Accuracy and validation of the model

Since the data is unbalanced, the accuracy for each trip purpose must be evaluated. The validation metrics precision and recall are used for this (see Table 1). *Precision* indicates how the predicted trip purposes correspond to the actual trip purposes. For example, 86% of the predicted trip purpose ‘home’ is actual the trip purpose ‘home’. *Recall* indicates how the actual trip purposes correspond to the predicted trip purposes. For example, 93% of the actual trip purpose ‘home’ is also the predicted trip purpose ‘home’. *F1-Score* is the harmonic average as a combination of both metrics $(2 * (Precision * Recall)/(Precision + Recall))$. Due to the harmonic average, low values are weighted more (*21*). This becomes evident by looking at the trip purpose ‘business’. While precision corresponds to 44%, recall only has a value of 8%. This means that the F1-Score only has 13%. Frequency corresponds to the sample size differentiated by trip purpose. Overall, accuracy meets 73%. Marco average is the average accuracy of the related metric. Weighted average is the average of the respective metrics weighted by the sample size of each trip purpose. Trip purposes ‘home’ (0.90) and ‘work’ (0.80) can be well identified by the trained model. This has already been made clear by the visualization (Figure 2 and Figure 3). Furthermore, more than half of the trip purposes of ‘shopping/errands’ (0.61) and ‘leisure’ (0.59) can be identified. Only 13% of trip purpose ‘business’ and 37% of trip purpose ‘pick up and drop off’ can be identified. This is mainly due to the fact that these trip purposes do not follow any time (standing time, start time) patterns, which makes the identification of the trip purposes more difficult. Taking into account that the model estimates trip purposes at trip level and no regularities have been considered in the model generation, an overall accuracy of 73% is quite acceptable.

**Table 1 Validation of the model for identification of the trip purpose**

| Trip purposes | Precision | Recall | F1-Score | Frequency |
|---|---|---|---|---|
| Home | 0.86 | 0.93 | 0.90 | 600,599 |
| Work | 0.80 | 0.80 | 0.80 | 191,847 |
| Business | 0.44 | 0.08 | 0.13 | 60,551 |
| Shopping/ errands | 0.55 | 0.67 | 0.61 | 275,147 |
| Leisure | 0.61 | 0.58 | 0.59 | 207,565 |
| Pick up and drop off | 0.47 | 0.30 | 0.37 | 115,855 |
| Accuracy | | | 73% | 1,451,564 |
| Macro average | 0.62 | 0.56 | 57% | 1,451,564 |
| Weighted average | 0.71 | 0.73 | 71% | 1,451,564 |

## RESULTS

The applicability of the CUMILE model, which is based on travel survey data, for identifying trip purposes in sensor data was shown in the previous section. We applied the trained model to the 7,489,686 trips captured in the sensor and estimated the purpose for each trip. In this section, we describe in the first part the distribution of the identified trip purposes in the sensor data. In the second part, we analyze differences in car usage between countries and between cars with different fuel types.

### Distribution of trip purposes in the sensor data

Figure 5 provides a comparison between the distribution of trip purposes in the CUMILE data and the identified trip purposes in the sensor data. We see that the distribution is similar. 'Business', 'leisure' and 'shopping/ errands' were estimated at a different share compared to the CUMILE data. As we mentioned before, this may be due to the variability of these activities, which cannot be described by characteristics such as the starting time. Nevertheless, the distribution of trip purposes in both data sets is of the same extent and deliver sound results for further analyses.

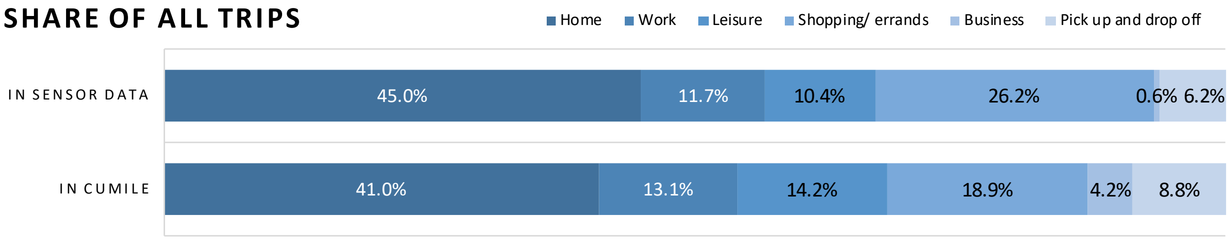

**Figure 5 Comparison of estimated trip purposes in sensor data and trip purposes in CUMILE**

Since the data set with the sensor data contains a large number of observations, the determination of the trip purposes allows to carry out more in-depth analyses with regard to car usage. When calculating the average distances distinguished by trip purpose in the sensor data differences become apparent. The average distance for 'work' is 11.8 km. Trips for 'shopping/

errands' are on average 6.3 km, trips with the purpose 'leisure' have an average distance of 7.9 km. The most frequent trip purpose 'home' has an average distance of 7.6 km.

### International Comparison

In the following, we investigate how car usage differs in the nine countries where the sensor data were collected. Table 2 gives an overview of car usage characteristics differentiated by the nine countries. It becomes clear, that in all nine countries included in the study, 'home' is the dominant trip purpose. In all countries, trips to home explain 44 to 46% of the trips. The car usage behaviors are also similar for the other trip purposes. 'Work' accounts for about 10% of all trips, while 'business' and 'pick up and drop off' are relatively rare trip purposes. An interesting result appears in the trip purpose 'shopping/ errands'. This category differs most between countries with a difference of up to 7.5 percentage points. In Austria 29% of all trips are errands or shopping trips. In Spain only about 22% of the trips are made by car for this reason. The comparatively few trips to go shopping in Spain are offset by a comparatively high share of trips for leisure activities (13%). In contrast, in Austria, only 9% of the trips in the sensor data were estimated as 'leisure'. Overall, differences between the countries are identified, a further assessment and interpretation will not be given in this study and is part of further research.

**Table 2 Characteristics of car trips by purpose differentiated by countries**

| | **Country** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Austria | Belgium | Germany | Denmark | Spain | France | UK | Italy | Nether-lands |
| **Characteristic** | | | | | | | | | |
| *Distribution of the trips across the countries in %* | | | | | | | | | |
| Share of trips across countries | 10.4 | 7.6 | 36.9 | 1.8 | 3.2 | 7.1 | 19.5 | 6.5 | 7.1 |
| *Distribution of trip purposes within country in %* | | | | | | | | | |
| Work | 10.2 | 11.2 | 12.2 | 12.1 | 12.8 | 12.1 | 11.1 | 11.4 | 12.1 |
| Business | 0.9 | 0.6 | 0.6 | 0.5 | 0.5 | 0.6 | 0.3 | 0.7 | 0.6 |
| Shopping/ errands | 29.1 | 27.2 | 26.1 | 27.3 | 21.6 | 26.1 | 25.8 | 24.7 | 25.6 |
| Leisure | 9.1 | 10.3 | 10.0 | 9.4 | 12.7 | 9.8 | 11.5 | 10.5 | 11.1 |
| Pick up and drop off | 6.7 | 6.3 | 5.9 | 6.7 | 6.2 | 6.2 | 6.2 | 7.8 | 5.8 |
| Home | 44.0 | 44.5 | 45.2 | 44.0 | 46.1 | 45.2 | 45.1 | 45.0 | 44.7 |
| Total | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| *Median distance of trips with purpose 'work' in kilometer by fuel type* | | | | | | | | | |
| Overall | 8.80 | 14.55 | 11.20 | 15.80 | 11.70 | 11.20 | 13.95 | 8.15 | 18.30 |
| Gasoline | 8.10 | 19.40 | 10.40 | 11.70 | 10.00 | 9.55 | 12.35 | 6.85 | 17.05 |
| Gasoline PHEV | 9.45 | 11.70 | 12.10 | 14.05 | 14.85 | 9.25 | 17.45 | 7.30 | 18.55 |
| Diesel | 8.95 | 13.10 | 11.60 | 23.90 | 12.25 | 12.05 | 14.35 | 8.40 | 27.60 |
| Diesel PHEV | 19.40 | 16.75 | 14.95 | 24.30 | 27.47 | 9.40 | 12.33 | 10.80 | 15.10 |

When examining the average distances per trip purpose, strong differences between the nine countries become visible. Purposes like 'leisure' and 'shopping' are more flexible in terms of time and location, which means that there is also a certain degree of flexibility in how individuals can fulfil these travel purposes. In the following we focus on the purpose 'work', as people are obligated to carry out such trips. In the sensor data in Italy, we see the lowest median distance to 'work' with 8.15 km. In Austria, the value of 8.8 km is also comparatively low. In contrast, Belgium and Denmark show high values with 14.55 km and 15.80 km. The Netherlands have the highest value with 18.3 km. This may be due to the spatial structure and infrastructure. On the one hand, other means of transport, such as the bicycle, may be used for commuting short distances in these countries instead of the car. On the other hand, longer distances in the car usage could be explained by a widespread spatial structure in rural areas.

Further insights are gained by including fuel type as car characteristic into the analysis. The sensor data includes not only conventional vehicles but also Plug-in Hybrid Electric Vehicles (PHEV). Regarding the distribution of the trip purposes differentiated by fuel type of the cars, no major differences are observable. However, when evaluating the median trip distance for commuting, we see differences (see Table 2). The examination of the standard deviations provides no further information. The median trip distances for the trip purpose 'work' vary between the fuel types. Diesel PHEV show for many countries the longest distances. This is of special interest, as the values exceed the average distances travelled by conventional diesel cars. This indicates that diesel PHEV are used in these countries, especially if the drivers are heavy users and tend to be commute longer distances. Exceptions are the Netherlands, U.K. and France. However, no general conclusion can be drawn as the number of diesel PHEV in the sensor data is considerably smaller than the number of the other fuel types. Across the countries, diesel PHEV explain only about 1% or less of the commuting trips. The comparison of the median commuting distances between conventional gasoline and diesel cars reveal only small differences with the tendency of slightly longer distances for diesel cars. In the Netherlands and Denmark there are clear differences in fuel types. In these countries, diesel cars are clearly used for longer trips to work. It is noticeable that in Belgium the situation is vice versa. Gasoline cars are used for longer distances than the diesel car. The comparison between conventional gasoline cars and gasoline PHEV also shows the tendency for the PHEV to be used for longer trips to 'work'. Gasoline PHEV are present to varying degrees in the data but explain between 1% (Italy) and 16% (Belgium) of trips to work. The fact that the two PHEV types are each used for in average longer distances to work than the conventional ones is an important finding, which is also relevant for understanding the market segments. This knowledge about the use of PHEV is helpful to develop further offers and measures to encourage more commuters to use PHEV. It has been shown that this type of vehicle is suitable for longer commuting distances across countries.

## LIMITATIONS

In the methodology presented, assumptions have been made, some of which are limitations of the approach. It must be considered that the data from CUMILE were used to learn the model. CUMILE is a unique data set as it describes car usage over one year, but it is based on survey data from Germany. Transferring the car usage model to the other European countries can suppress country-specific characteristics (e.g., siesta in Spain). In principle, a good level of comparability exists, as is shown by the average annual mileage of passenger cars in a European comparison (*22*). In addition, since we assume no major discrepancies, the usability for other European countries seems appropriate. When assessing the results, it should be mentioned again that the collected sensor data, includes only premium cars, whereas CUMILE contains all types of cars. A conclusion of representative statements for the car fleet is therefore not possible.

There are also limits regarding the estimation of trip purposes. Since the generated model estimates on trip level, we did not consider any regularity in car use. Trips to work for example are characterized by a higher repetition than trips for leisure purposes. Further model extensions could also take into account aspects that reach beyond a single trip.

## CONCLUSIONS

In our study, we deal with the issue that in sensor data there is no information about the motives and purposes of car usage. Using a large set of sensor data and car usage data from a model based on survey data, we identified the purpose of the trips in the sensor data. For this, a model was trained on the car usage data and then applied on the sensor data to estimate the specific trip purposes. Input variables were standing time and start time of the trips, which proved to be important variables to describe characteristics of specific trip purposes. The approach allows a differentiated analysis of car usage and to understand why and for what reasons people use cars. Information that can only be obtained from travel survey data is analyzed without the disadvantage that information on the cars used is missing. It brings together the advantages of two different types of data and surveys without requiring participants to validate the data collected.

The results show that the distribution of the trip purposes and especially the trips with the purposes ‘home’ and ‘work’ show only little differences across the nine analyzed countries. This confirms that these trip purposes can be well determined with the selected attributes. However, we also see that travel purposes that vary greatly in terms of start time and standing time, such as ‘shopping/ errands’, also vary across the countries. Further differences between the countries were identified with regard to the distances of trips to ‘work’ identified in the sensor data. Belgium, Denmark and the Netherlands show the longest commuting distances in the car usage. This indicates that transport alternatives are used for shorter commuting distances in these countries. However, for longer commuting distances, the car is the chosen means of transport. Using the advantage of having detailed information about the cars in the sensor data, we identified different usage characteristics of the fuel types. PHEVs have a longer commuting distance than their conventional counterparts in almost all countries. In some countries this may be due to the fact that the purchase of PHEVs is supported by tax benefits.

Overall, we see similarities and differences between the countries with the sensor data. This can be measured directly and uniformly because the data collection with the sensors and the generated model to identify trip purposes was the same in all countries. In contrast to data from travel surveys, where each country follows a different survey approach, the sensor data is directly comparable and allows comparisons between the car usages in the countries. We emphasize that the differentiation of car usage according to car characteristics plays a special role and that sensor data allows such an investigation. Especially for the development of measures and market-oriented offers to influence the car use behavior of individuals the understanding of usage patterns is highly relevant.

For further research, we suggest including further information from the sensor data in the identification process of the trip purposes, such as seat occupation. It is likely that the seat occupancy is higher for leisure and pick up and drop off than for example for commuting to work. Further, more research is needed to describe country-specific characteristics. By analyzing the different countries, we could see that the distribution of trip purposes is similar. An analysis of specific areas (e.g. rural or urban areas) where the car is primarily located would provide further insides. It would then be possible to quantify the extent to which, for example, the use of cars in urban areas varies from country to country.

## AUTHOR CONTRIBUTIONS

The authors confirm contribution to the paper as follows: study conception and design: U. Niklas, S. von Behren; literature review: U. Niklas, M. Magdolen; data preparation: U. Niklas; data analysis: U. Niklas; interpretation of results: U. Niklas, M. Magdolen, S. von Behren, P. Vortisch; draft manuscript preparation: M. Magdolen, U. Niklas. All authors reviewed the results and approved the final version of the manuscript.

## REFERENCES

1. Alexander, L., S. Jiang, M. Murga, and M. C. González. *Origin–destination trips by purpose and time of day inferred from mobile phone data. Transportation Research Part C: Emerging Technologies*, Vol. 58, 2015, pp. 240–250, http://dx.doi.org/10.1016/j.trc.2015.02.018.
2. Calabrese, F., M. Diao, G. Di Lorenzo, J. Ferreira, and C. Ratti. *Understanding individual mobility patterns from urban sensing data: A mobile phone trace example. Transportation Research Part C: Emerging Technologies*, Vol. 26, 2013, pp. 301–313, http://dx.doi.org/10.1016/j.trc.2012.09.009.
3. Isaacman, S., R. Becker, R. Cáceres, S. Kobourov, M. Martonosi, J. Rowland, and A. Varshavsky. *Identifying Important Places in People's Lives from Cellular Network Data. In Pervasive computing. 9th international conference, Pervasive 2011, San Francisco, USA, June 12 - 15, 2011 ; proceedings*, K. Lyons, J. Hightower and E.M. Huang, eds. Springer, Berlin, 2011, pp. 133–151.
4. Phithakkitnukoon, S., T. Horanont, G. Di Lorenzo, R. Shibasaki, and C. Ratti. *Activity-Aware Map: Identifying Human Daily Activity Pattern Using Mobile Phone Data. In Human Behavior Understanding*, A.A. Salah, T. Gevers, N. Sebe and A. Vinciarelli, eds. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010, pp. 14–25, http://dx.doi.org/10.1007/978-3-642-14715-9_3.
5. Gong, L., T. Morikawa, T. Yamamoto, and H. Sato. *Deriving Personal Trip Data from GPS Data: A Literature Review on the Existing Methodologies. Procedia - Social and Behavioral Sciences*, Vol. 138, 2014, pp. 557–565, http://dx.doi.org/10.1016/j.sbspro.2014.07.239.
6. Bohte, W., and K. Maat. *Deriving and validating trip purposes and travel modes for multi-day GPS-based travel surveys: A large-scale application in the Netherlands. Transportation Research Part C: Emerging Technologies*, Vol. 17, No. 3, 2009, pp. 285–297, http://dx.doi.org/10.1016/j.trc.2008.11.004.
7. Seo, T., T. Kusakabe, H. Gotoh, and Y. Asakura. *Interactive online machine learning approach for activity-travel survey. Transportation Research Part B: Methodological*, Vol. 123, 2019, pp. 362–373, http://dx.doi.org/10.1016/j.trb.2017.11.009.
8. Xiao, G., Z. Juan, and C. Zhang. *Detecting trip purposes from smartphone-based travel surveys with artificial neural networks and particle swarm optimization. Transportation Research Part C: Emerging Technologies*, Vol. 71, 2016, pp. 447–463, http://dx.doi.org/10.1016/j.trc.2016.08.008.

9. Westat. *2017 NHTS Data User Guide*, 2018.

10. Ecke, L., B. Chlond, M. Magdolen, T. Hilgert, and P. Vortisch. *Deutsches Mobilitätspanel (MOP) - Wissenschaftliche Begleitung und Auswertungen, Bericht 2018/2019: Alltagsmobilität und Fahrleistung,* Institut für Verkehrswesen (KIT), 2020.

11. Leerkamp, B. *Bundesweite Verkehrsverflechtung 2015 im Motorisierten Individualverkehr (MIV) und Schwerverkehr (SV) - Analysen auf Basis satellitengestützter Daten,* Berlin, 28./29.11.2019.

12. infas, DLR, IVT Research, and infas 360. *Mobilität in Deutschland - Ergebnisbericht*, 2017.

13. Eisenmann, C. *Mikroskopische Abbildung von Pkw-Nutzungsprofilen im Längsschnitt*, 2018, http://dx.doi.org/10.5445/IR/1000085513.

14. Eisenmann, C., and R. Buehler. *Are cars used differently in Germany than in California? Findings from annual car-use profiles. Journal of Transport Geography*, Vol. 69, 2018, pp. 171–180, http://dx.doi.org/10.1016/j.jtrangeo.2018.04.022.

15. Niklas, U., S. von Behren, B. Chlond, and P. Vortisch. *Electric Factor—A Comparison of Car Usage Profiles of Electric and Conventional Vehicles by a Probabilistic Approach. World Electric Vehicle Journal*, Vol. 11, No. 2, 2020, p. 36, http://dx.doi.org/10.3390/wevj11020036.

16. Chlond, B., C. Weiss, M. Heilig, and P. Vortisch. *Hybrid Modeling Approach of Car Uses in Germany on Basis of Empirical Data with Different Granularities. Transportation Research Record: Journal of the Transportation Research Board, No. 2412*, 2014, pp. 67–74, http://dx.doi.org/10.3141/2412-08.

17. Zumkeller, D., B. Chlond, J. Last, and W. Manz. *Long-Distance Travel in a Longitudinal Perspective: The INVERMO Approach in Germany. TRB 85th Annual Meeting Compendium of Papers*, Washington, D.C., 2006.

18. Niklas, U., S. von Behren, C. Eisenmann, B. Chlond, and P. Vortisch. *Premium factor – Analyzing usage of premium cars compared to conventional cars. Research in Transportation Business & Management*, 2020, http://dx.doi.org/10.1016/j.rtbm.2020.100456.

19. Mokhtari, A., and A. Ribeiro. *Global convergence of online limited memory BFGS. 16.1 (2015). The Journal of Machine Learning Research*, Vol. 16, No. 1, 2015, pp. 3151–3181.

20. Friedman, J., T. Hastie, and R. Tibshirani. *The elements of statistical learning.* Springer series in statistics, New York, 2001.

21. Géron, A. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow.* O'Reilly Media, Inc, 2019.

22. Odyssee-Mure. *Sectoral Profile - Transport*, 2020. https://www.odyssee-mure.eu/publications/efficiency-by-sector/transport/transport-eu.pdf.