APPROACHES TO CREATE A DATA BASIS FOR MODELLING LONG-DISTANCE TRAVEL BEHAVIOUR

Ursula Pfefferkorn
German Aerospace Center (DLR), Swiss Federal Institute of Technology in Zurich (ETHZ)

1. INTRODUCTION

Long-distance travel, respectively interurban travel, makes up a considerable share of transport performance: trips of more than 100 km distance account for nearly half of the passenger transport performance. Furthermore, passenger long-distance travel is still characterized by considerable growth while growth in other segments has slowed down (Manz, 2005). Moreover, passenger long-distance travel is in the focus of politics and research against the background of energy and climate objectives and due to a dynamic market situation, for example because new providers enter the market (long-distance busses) or because of continuous price competition (air carriers).

Despite the relevance of this transport segment, long-distance travel is often not adequately represented in transport models. In Germany one fundamental reason for that is the heterogeneous data situation, resulting by the different interests of the institutions that generate relevant data. The question arises whether it is possible to generate a consistent and comprehensive data set from the different data sources, which could be a basis for a long-distance transport model. The paper aims to introduce the complexity of the long-distance data landscape in Germany and questions whether the currently available data is sufficient to represent long-distance travel behaviour.

The issues above are addressed in the current dissertation research project started in January 2016. The results of the first phase of the project are briefly summarized in the present paper. In a first step, the paper intends to clearly differentiate and define the segment of long-distance travel. In the second step the available data sources are listed and compared with each other with the aim to approximate the realistic boundaries for demand figures of long-distance travel. Finally the paper presents a methodological framework on how data that is currently available to modellers can be combined to generate a comprehensive current data basis for modelling long-distance travel behaviour.

2. LONG-DISTANCE TRAVEL

2.1 Characteristics and available data sources

While long-distance travel represents only a very small proportion of trips (about 1.3 % of trips within Germany, cf. infas & DLR, 2010, own calculation), it accounts for a substantial part of the mileage: according to Frick and Grimm (2014) about 45 % of the total mileage travelled can be ascribed to long-distance travel. Long-distance travel hereby contains of all trips including international ones which exceed a distance of 100 km. However, it is important to note that long-distance travel demand is very unevenly spread across the population: Manz (2005) stated that about 50 % of the population in Germany accounts for 90 % of all long-distance trips.

Although widely-used in the field of transportation research, the term “long-distance travel” is not standardized. Given that different institutions (e. g. public authority, private enterprises, and research institutes) are interested in the field of long-distance travel for various purposes, the delimitations are not congruent. Since the travel distances are normally generated from surveys the information that can be captured by the resulting data strongly depends on the applied survey methodology. Already (Kuhnimhof, Frick, Grimm, and Phleps (2014)) stated, that several data collections differ according to the following survey focusses:

- Travel purpose, i.e. personal and business travel,
- Duration, i.e. trips with and without overnight stay, and
- Touristic travel (involving travel outside one’s usual environment) and everyday travel (travel within one’s usual environment).

For this paper the objective is a compilation and comparison of different data sources with focus on long-distance travel. A common ground for defining long-distance travel as a whole lies in a distance-based classification, since the distance of trips is always captured in the surveys. Therefore, I hereby chose a threshold of 100 km to define long-distance travel according to most previous studies on long-distance travel.

A large amount of data which are directly or indirectly related to travel behaviour is collected in different ways and for different purposes. If the data is collected for commercial purposes (e. g. market research) or in a passive way (e. g. mobile phone data) then the access to the data is either very expensive (since it is dedicated to business companies, not research purposes) or not possible due to privacy regulations. Some data is available for research purposes because it is either surveyed on behalf of the public authority (e. g.

MiD, MOP) or because the fee was paid in the context of a research project (e. g. Reiseanalyse). The ideal case is when the data is available on the level of individual respondents, also called “micro-level”. However, survey results are often only published as aggregate figures, for instance in tables or graphs. Consequently, only part of the information is available, so the analysis of the interaction between several variables or the separate analysis of a certain subset of people or trips is impossible.

The following section presents data with information on long-distance travel that is considered for use for the generation of the long-distance data set.

a) Mobilität in Deutschland („MiD“)

MiD (Mobility in Germany) is a recurring Germany-wide travel survey which primarily aims to obtain data on households’ everyday mobility. The current survey dates back to 2008. An update is scheduled for this year, 2016. On behalf of the German BMVI (Federal Ministry of Transport and Digital Infrastructure), the Institute for Applied Social Science Research (infas) and German Institute of Economic Research (DIW) developed the design of the survey and carried it out. For the MiD 2008, the German Aerospace Center (DLR) also participated in the survey process. The survey mainly consists of two modules: first a one-day trip diary and second a journey part with retrospective questions on the three most recent overnight trips of the past three months. It is highly relevant for transport research because compared to other transportation surveys in Germany it has a much larger sample size of about 50,000 households and 100,000 persons. There are two microdata sets available: MiD 2008 and MiD 2002.

b) Deutsches Mobilitätspanel (“MOP”)

The German Mobility Panel (MOP) surveys travel demand and personal mobility behaviour in Germany on an annual basis. The study is conducted, similar as the MiD, on behalf of the BMVI. It is designed as a panel, where households participate in three consecutive years. On a household level, persons are asked to report their mobility behaviour for the period of one week. A disadvantage of the study is that it covers only certain weeks of the year which are deliberately outside holiday seasons since the focus of the survey is everyday mobility. The sample size in 2012/2013 was about 1,200 households with 2,400 persons. A microdata set of the MOP is available for the period between 1994 and 2014.

c) Reiseanalyse ("RA")

The "Reiseanalyse" is an annual population-representative survey which aims to collect and describe holiday and journey behaviour as well as holiday motivations and interests of the German-speaking population in Germany. The RA is conducted by the „Forschungsgemeinschaft Urlaub und Reisen e. V." (FUR). The survey includes holiday trips with duration of more than five days and short holiday trips of two to four days. The here available microdata date back to 2011. The sample size of this data set is about 9,400 persons.

d) dwif Tagesreisen der Deutschen („DWIF")

This survey was executed by dwif e. V. and jointly financed by the German BMWi (Federal Ministry of Economics and Technology), the German federal states, the ADAC (General German Automobile Club) and the dwif Consulting GmbH. The representative sample of the survey consists of 36,400 people interviewed by telephone between May 2012 and April 2013. In the study all trips occurring outside one's daily surrounding were defined as "day trips". For this survey, only aggregated data from the final report is available.

e) VDR Geschäftsreiseanalyse ("VDR")

Conducted by the "Verband Deutsches Reisemanagement e. V. (VDR)", the Geschäftsreiseanalyse study aims to survey, analyse and describe the business travel behaviour of German enterprises representatively on an annual basis. A random sample of about 800 enterprises registered in Germany and with at least 10 employers was interviewed in the latest study of 2016. Only a final report with tables and operating figures is available. Unfortunately, within the results no distance-based differentiation is made (cf. VDR Verband Deutsches Reisemanagement e.V. (2016)). An important difference to the earlier mentioned surveys is that this is not a household or person survey, but a company survey, meaning that one responsible person of a company speaks for the group of business travelling employees. Therefore the socio-demographic characteristics of the single persons are not known.

f) Geschäftsreisendenstudie Bad Honnef ("BH")

The International University of Applied Sciences Bad Honnef and the Institute for Applied Social Science Research (infas) realised a study on business travel behaviour in 2009. A population representative sample of 600 business travellers was interviewed and only journeys of a minimum distance of 50 km were captured. The results are available only as aggregates in a final report (cf. Schneider (2009)).

2.2 Establishing realistic figures for long-distance trip frequencies

When aiming to create a data basis for long-distance travel behaviour one crucial component is the reproduction of correct marginal totals over the population. One such target value is the average number of long-distance trips per capita and year.

In former research, there are two approaches for quantification of long-distance travel which will be labelled in this paper as the “universal approach” and the “consolidation approach”. The “universal approach” is based on one certain data set which covers all trip purposes and the “consolidation approach” bases upon different sources that cover only certain segments and therefore must be combined.

“Universal approach”

Within the “universal approach”, there are two different kinds of data usable to draw conclusions on long-distance travel characteristics. One possibility to estimate an average number of long-distance trips per capita and year provide household travel surveys like MiD and MOP, whereat the journey data set of MiD is not usable for this universal approach, since it covers only overnight journeys. Both MiD 24h travel diary and MOP use a diary format to ask persons and households for all trips made during one random day (MiD) or one week (MOP). The amount of long distance trips is determined by extrapolating all long-distance trips over one year. However, since the reporting period is very short, this procedure bears some difficulties when focussing only on trips greater than 100 km:

- The sample size of persons who made trips longer than 100 km is rather small; the representativeness is questionable.
- People report each stage of a journey and therefore single trips must not exceed the 100 km distance although the trip chain as a whole would be experienced as a long-distance trip.
- In the MiD, people are asked to report the past day, so it can be assumed that all people who have gone on journeys with more than one overnight stay on that day are not detected completely.

The second data type is collected with special focus on long-distance travel, which was the case in the projects INVERMO (Zumkeller, Manz, Last, & Chlond, 2005), DATELINE (Brög, Erl, & Schulze, 2003), and KITE (Wirtz, Zumkeller, Chlond, & Schlosser, 2008). Those projects aimed to circumvent the difficulties mentioned above by considering long-distance travel detached from everyday travel already within the survey methodology. Such kind of surveys usually asks for long-distance trips retrospectively, in which the outcome quality regarding realism and comparability is strongly affected by a number of aspects:

- Response burden (e. g. participation in several interview waves, high time effort, fatigue effects) (cf. K W Axhausen, Schmid, & Weis, 2015)
- Length of retrospective reporting period (memory effects)
- Interview method (written: paper/online or oral: personal, telephone)
- Survey season, if not the whole year is surveyed
- Sample size and population representativeness

Considering these points one can assume that trip frequencies are too low when estimated using the "universal approach".

"Consolidation approach"

Another way of estimating the long-distance trip frequency is to subdivide long-distance travel into separate segments, here referred to the so-called "consolidation approach". The objective of this approach is to calculate the trip frequency as a sum of single segments without double-conting of trips due to overlapping of segments. This is motivated by the fact that several studies and data collections only cover certain parts of long-distance travel, like holiday journeys (Reiseanalyse, RA), business trips (VDR Geschäftsreiseanalyse; Geschäftsreisendenstudie), overnight trips (MiD journey data set) or day trips (dwif Tagesreisen der Deutschen). Frick and Grimm (2014) followed this approach within their research project on long-distance mobility.

In Table 1 annual long-distance trip frequencies from relevant data sources are summarized. Long-distance commuting is hereby included. The objective of the compilation is to derive a range of long-distance trip frequencies which is certainly not exceeded. In the next step the ranges within different segments of long-distance travel can also be determined. That is one requirement for calibrating an incomplete data set later on.

The upper part of the table (lines 1-5) lists findings where for only the sum of the annual trip frequency is stated, followed by the two studies INVERMO and DATELINE, where trip rates for different trip purposes and the sums are known. The next three lines list studies that only treat special trip purposes and therefore summation is impossible. The ifmo study is mentioned the very last line because it differs from the other studies. Like in the present study, it combines different data sources and therefore serves only for evaluation.

Trip frequency = Number of trips ≥ 100 km per person and year

Survey	Holiday trips (5+ days)	Short holiday trips (2-4 days)	Other personal overnight trips	Personal day trips	Private everyday trips	Long distance commuting	Business everyday trips	Overnight Business trips	Business day trips	Sum Σ	Source
MiD 2008 (1-day data set)	▨	▨	▨	▨	▨	▨	▨	▨	▨	**8.3**	MiD 2008, Basisdatensatz (own calc.)
MiD 2002	▨	▨	▨	▨	▨	▨	▨	▨	▨	**8.5-9.0**	Zumkeller, Manz, Last, and Chlond (2005)
MOP 2013	▨	▨	▨	▨	▨	▨	▨	▨	▨	**10.6**	MOP 2013, Basisdatensatz (own calc.)
MOP 2002	▨	▨	▨	▨	▨	▨	▨	▨	▨	**8.4-9.0**	Zumkeller, Manz, Last, and Chlond (2005)
KITE 2008-09 (country average)	▨	▨	▨	▨	▨	▨	▨	▨	▨	**10.1**	Frei, Kuhnimhof, and Axhausen (2010)
INVERMO 2001	1.6	4.6				1.3	1.3			**8.8**	Zumkeller, Manz, Last, and Chlond (2005)
DATELINE 2001 + MiD 2002	1.4		5.0				1.9			**8.3**	Hautzinger, Stock, and Schmidt (2005)
MiD 2008 (journey data set)	1.9	1.5	0.8	▨	▨	▨	▨	0.7	▨	▨	MiD 2008, Basisdatensatz (own calc.)
RA 2016	1.3	2.4	▨	▨	▨	▨	▨	▨	▨	▨	Sonntag and Schrader (2016)
dwif Tagesreisen der Deutschen	▨	▨	▨	6.0	▨	▨	▨	▨	2.5	▨	Harrer, Zeiner, Maschke, and Scherr (2013)
IPK International: Deutscher Reisemonitor 2010 *[1]	2.0		1.8	▨	▨	▨	▨	0.7	▨	▨	DZT (2013)
VDR Geschäftsreiseanalyse *[1]	▨	▨	▨	▨	▨	▨	▨	1.0	1.3	▨	VDR Verband Deutsches Reisemanagement e.V. (2016)
FH Bad Honnef (BH) Geschäftsreisende 2009 *[2]	▨	▨	▨	▨	▨	▨	▨	1.5		▨	Schneider (2009)
a) Min. value per segment	1.3	1.5	0.8	6.0	-	1.3	-	0.7	1.3	**12.9**	
b) Max. value per segment	1.9	2.4	1.8	6.0	-	1.3	-	1.0	2.5	**16.9**	
c) Ifmo Langstreckenmobilität 2014	1.0	1.2	0.3	6.0	1.0	2.0	2.0	1.2	1.2	**15.9**	Frick and Grimm (2014), various data sources

*[1] No minimum distance

*[2] Trips ≥ 50 km

Table 1: Consolidation of long-distance trip frequencies from different sources

The lowest *total* number of trip frequencies found in recent studies was 8.3 long-distance trips per capita and year, estimated within a project that combined the DATELINE data with day trip data from MiD 2002 (cf.Hautzinger, Stock, & Schmidt, 2005). Since daytrips where underrepresented in the original DATELINE data it was enhanced by the MiD data on day trips. This value value will stand for the lower limit for the certain range of long-distance trip frequency as can be seen in Figure 1.

To infer the lower and upper limits for the “supposable range” of long-distance trip frequencies I took the lowest and highest mentioned values per segment and summed them up, which yielded a trip frequency of 12.9 for the lower limit and 16.9 for the higher limit as can be seen in Table 1 line a) and b), as well as in Figure 1.

This was exactly the approach in the ifmo project also (Frick & Grimm, 2014) which is mentioned in the lower part of Table 1, line c). The result of the ifmo study lies within the inferred “supposable range” although the values of the single segments differ. Remarkably the upper limit of the supposable range is around twice as high as within calculations using the “universal approach” (8.3 up to 10.1 trips per capita and year). Overlapping cannot be ruled out within the approach of estimating the trip frequency from single segments at the moment. This is because the value of 16.9 is set as the upper limit of the supposable range of long-distance trip frequency and it seems very unlikely that the real value is beyond this range. When long-distance commuting is excluded the upper limit of the range is 15.6.

The range is illustrated in Figure 1.

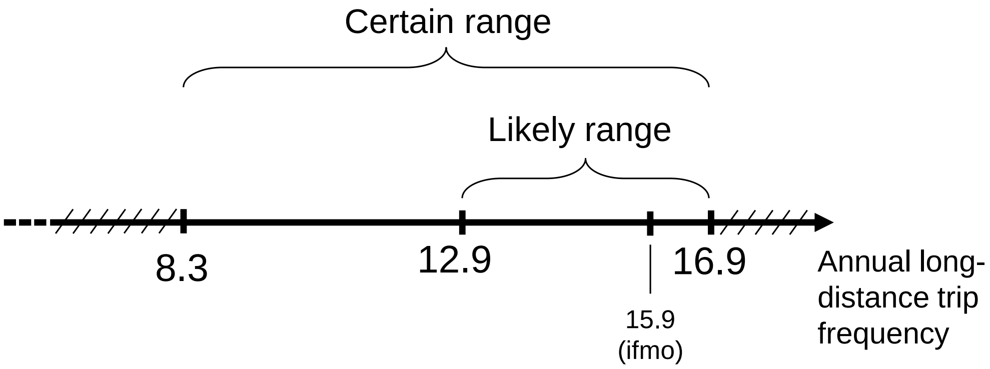

Figure 1: Certain and supposable range of long-distance trip frequency

3. ENHANCING THE CURRENT DATA BASIS

3.1 Approach

Transport models are an essential tool to estimate the impact of measures on the travel behaviour. Since people’s transport decisions are very complex, this is usually done using transport models. For the quality of the model results it is crucial that the data used to model the decision behaviour is representative for the whole population. Contrary to the long-distance travel segment, every-day travel behaviour can be modelled well by a date survey of a sufficient sample of the population. It is supposed that long-distance travel is inadequately represented in a 24h travel diary survey like MiD because the number of cases gets small when only long-distance trips are considered. The general concept for the present research project is to enrich the existing data basis on long-distance travel by using several data sources and combining them to a single data set. The aim is to generate a microscopic data set which acts like a fictive long-distance travel behaviour survey over a whole year. This data set then enables the calculation of trip rates and mileage for different groups of the population.

Figure 2 illustrates the course of action to create a micro data file on long-distance travel. The core of this approach is the initial separate consideration of overnight trips and day trips to create two microscopic data files that represent certain known marginal sums. The different data sets within the two segments “overnight trips” and “day trips” will be combined (“Data Fusion”). After having two representative data sets for both overnight trips and day trips both data sets shall be first calibrated with known marginal distributions (“Calibration”) and afterwards combined to one long-distance travel data set (“Consolidation”). This will finally be evaluated with earlier works on the quantitative analysis of long-distance travel and their inner distributions of characteristics (“Evaluation”).

In the following sections I will describe and justify the single method parts of the process illustrated in Figure 2 more in detail.

3.2 Data fusion

Long-distance overnight trips

As summarized above, there are two data sources available on the micro level for the segment of overnight trips: One is the journey data set of the MiD and the other the data of the RA 2011 survey.

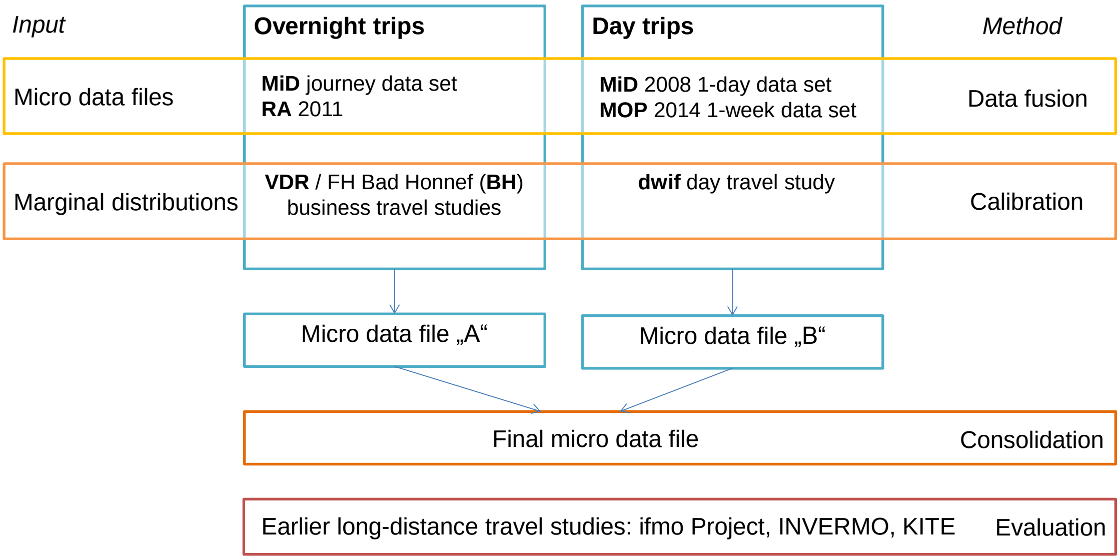

Figure 2: Methodological framework of creating a micro data file on long-distance travel

The journey data set is a sub-dataset of the MiD survey. The participants of the main study where asked how many overnight journeys they have made during the past three months. Those respondents whose number of overnight journeys during the past three months was greater than zero were then asked about the characteristics of their three latest overnight journeys. Each reported journey is one data set entry in the journey data set. The data set consists of 20,665 persons of 14,971 households who reported 36,182 overnight journeys in total. The overall sample size of the MiD is 60,713 persons of 25,922 households. This means that about every third person has done at least one overnight journey during the past three months.

The RA survey asks only for private holiday trips and has two parts. One is the holiday data set where the interviewees report all trips with four or more overnight stays within the past year. It contains of 7,694 persons and 7,648 holiday journeys which were made by 5,809 persons, meaning that about 75 % of all persons do at least one holiday journey with five or more overnight stays during one year. The other part is the data set of short holiday trips (one to three overnight stays) consisting of 2,829 journeys made by 1,793 persons, listing only persons with at least one short holiday trip. Slightly more than every second person makes at least one short holiday trip per year. The interviewees of both survey parts are not the same.

Both data sets have strengths and weaknesses. The RA data set is surely the best choice for long holiday trips since it covers the whole year and memory loss is expected to be little due to the length of the journeys. Furthermore, for all overnight trips that do *not* take place within in a private holiday setting only the MiD journey data set can give information on the micro level.

Within the segment of short holiday trips it requires diligent consideration of which of the both data sets are suitable. Both the journey data set of MiD as well as the RA data set have pros and cons: the first one only covers a time period of three months and a maximum of three journeys in detail. However, memory loss is expected to be smaller than in the RA survey, where the retrospective reporting period is six months. Though, in the RA survey a whole year is covered (2 six-month periods) and a maximum of five journeys can be reported. However, the sample size of the RA survey is much smaller than of the MiD journey data set.

Long-distance day trips

The only available micro data files to represent day trips are the MiD date dataset (one-day trip diary) and the MOP mobility panel (one-week trip diary).By applying several data queries under certain assumptions the long-distance day trips can be extracted.

For the MiD data file this has been done with the same procedure which was applied by Hautzinger et al. (2005), who provided an algorithm to extract all day trips from the MiD data set by looking only at closed trip chains. Of all people participating in the diary survey (over 51.000 persons) 623 persons have done a day trip with a distance of more than 100 km. Since this is rather a small sample size it will be difficult to draw conclusions on the basis of one year connected with characteristics of different person groups. Nevertheless, presently there is no better micro data on long-distance day trips available.

3.3 Calibration, Consolidation and Evaluation

Several studies related to the topic of long-distance travel only publish final reports with key numbers or summary tables but no micro data of the surveys. These results can still be used since they provide target numbers for the calibration of the raw data set. For the overnight segment an overnight travel study (DZT, 2013) and two business travel studies (Schneider, 2009; VDR Verband Deutsches Reisemanagement e.V., 2016) are significant supplements. For day trips the results of the study “Tagesreisen der Deutschen” (Harrer, Zeiner, Maschke, & Scherr, 2013) are a crucial amendment. After the calibration the micro data files should now represent the explored values of trip frequency and mileage within these reports.

The next step planed for the present research project is the consolidation of the two day trip and overnight micro data sets into one long-distance travel data set. Afterwards the data set can be analysed and compared with earlier studies like ifmo, INVERMO or KITE.

4. SUMMARY AND OUTLOOK

The present paper introduces the complexity of the long-distance travel data landscape in Germany. By listing earlier studies it illustrates the diversity of published indication of annual long-distance trip frequencies and sets a range for it. It can be assumed that the true value for an annual long-distance trip frequency supposedly lies somewhere between 13 and 17. A more precise specification is not possible presently which reveals today's huge uncertainties relating to the quantification of long-distance travel behaviour. It is shown that a consolidation approach using a combination of several data sources leads to higher long-distance trip frequencies as if they result from only one of the sources. Currently there is no micro-level data set that contains all aspects of long-distance travel comprehensively. However, several studies and surveys cover parts of long-distance travel in different forms of data availability. The huge challenge will be to create a combined data set that is free of overlapping of the single travel segments to avoid overestimation of long-distance trip frequencies. A methodological framework to create such data set is presented in this paper, while this idea is strongly reasoned by the here available data. Since overnight trips are considered separately in two important available micro-level data sets, this separation between overnight travel and day trips is adopted in the here presented approach. This must not hold for different conditions of data availability. In forthcoming work the methods of the single parts (data fusion, calibration, consolidation, and evaluation) have to be concretised and applied.

REFERENCES

Axhausen, K. W., et al. (2015). Predicting response rates updated. *Arbeitsberichte Verkehrs-und Raumplanung, 1063*.

Axhausen, K. W. and M. Youssefzadeh. (1999). MEST, Methods for European Survey of Travel Behavior: Project.

Brög, W., et al. (2003). DATELINE Design and Application of a Travel Survey for Long-distance Trips Based on an International Network of Expertise Concept and Methodology.

DZT. (2013). *Das Reiseverhalten der Deutschen im Inland*. Frankfurt a. M.: Deutsche Zentrale für Tourismus e.V. (DZT).

Frick, R. and B. Grimm. (2014). Langstreckenmobilität – Aktuelle Trends und Zukunftsperspektiven, Grundlagenstudie.

Harrer, B., et al. (2013). *Tagesreisen der Deutschen*. München: dwif e. V.

Hautzinger, H., et al. (2005). Erstellung von Microdatenfiles zu Ein-und Mehrtagesreisen auf Basis der Erhebungen MiD und DATELINE, Schlussbericht. *Institut für angewandte Verkehrs-und Tourismusforschung eV, Heilbronn/Mannheim*.

infas and DLR. (2010). *Mobilität in Deutschland 2008. Basisdatensatz. Survey on behalf of the German Federal Ministry of Transport, Building and Urban Development*.

Kuhnimhof, T., et al. (2014). *Long Distance Mobility in Central Europe: Status Quo and Current Trends.* Paper presented at the European Transport Conference 2014.

Manz, W. (2005). Mikroskopische längsschnittorientierte Abbildung des Personenfernverkehrs. *Schriftenreihe des Instituts für Verkehrswesen der Universität Karlsruhe (TH), 62/05*.

Schneider, J. (2009). *Geschäftsreisende 2009* Internationale Fachhochschule Bad Honnef / Bonn.

Sonntag, U. and R. Schrader. (2016) Reiseanalyse 2016. *Vol. 46. Erste ausgewählte Ergebnisse der Reiseanalyse*. Kiel: FUR Forschungsgemeinschaft Urlaub und Reisen e.V.

VDR Verband Deutsches Reisemanagement e.V. (2016) VDR Geschäftsreiseanalyse 2016. *Vol. 14*. Frankfurt a. M.

Wirtz, M., et al. (2008). *KITE - A Knowledge Base for Intermodal Travel in Europe.* Paper presented at the European Transport Conference 2008; Proceedings.

Zumkeller, D., et al. (2005). Die intermodale Vernetzung von Personenverkehrsmitteln unter Berücksichtigung der Nutzerbedürfnisse (INVERMO). *Schlussbericht, März*.