

# Determination of long-distance travel demand: How to merge distinct data sources into a meaningful and consistent picture?

Angelika SCHULZ\*<sup>1</sup>, Tobias KUHNIMHOF<sup>2</sup>, Miriam MAGDOLEN<sup>3</sup>, Bastian CHLOND<sup>3</sup> & Katja KÖHLER<sup>1</sup>

\* corresponding author: [angelika.schulz@dlr.de](mailto:angelika.schulz@dlr.de)

<sup>1</sup> German Aerospace Center (DLR) – Institute of Transport Research (Berlin, Germany)

<sup>2</sup> RWTH Aachen University – Institute of Urban and Transport Planning (Aachen, Germany)

<sup>3</sup> Karlsruhe Institute of Technology (KIT) – Institute for Transport Studies (Karlsruhe, Germany)

## Abstract

In contrast to everyday travel, long-distance passenger travel continues to increase for the German population resulting in a share of roughly 40 to 50% of total passenger kilometers. Although data are available for some travel segments (e.g., long-distance rail, air travel), there is no up-to-date and consistent overall picture of this specific demand, neither in terms of total transport volumes, nor in terms of socio-demographic characteristics of the corresponding population and the driving forces. A dedicated survey – already carried out in the early 2000s (INVERMO 2002/2003) – revealed that long-distance travel was extremely unevenly distributed among the population: only a small proportion was responsible for the majority of travel volumes. This small proportion was characterized by high education levels, higher incomes and also urban life styles. The question arises whether this fact has already changed in view of current societal processes, such as an increasing level of education, the conversion to a knowledge-based society, long-term socialization in terms of travelling to distant destinations, the emergence of multi-local lifestyles or wide-ranging social networks, all of them likely to result in changing travel volumes and shifting proportions between different groups of travelers.

Analyses in this respect, however, require a meaningful data framework. The challenge is to gather and prepare available data according to a suitable definition of 'long-distance travel' and to close data gaps based on complementary data collection or any other empirically founded assumptions.

Despite their focus on everyday mobility, traditional National Household Travel Surveys (NHTS) continue to be the main source of data for describing and quantifying long-distance travel. In the case of the German NHTS (MiD – Mobility in Germany), however, there are few aspects to consider when extracting long-distance trips. On the one hand, such trips are subject to underreporting: While respondents are asked to report actual trips made on a designated reporting day only, long-distance trips usually take place at lower frequencies. In addition, unit non-response is an issue, as highly active people responsible for a large share of long-distance trips are very likely to be absent during the respective survey period. To address these problems, a dedicated 'travel module' for respondents aged 14 and over aims to cover overnight journeys for the last three months regardless the distance. Due to the 'overnight stay' criterion, however, day trips, which can also cover longer distances, are not included. The same is true for journeys made by children below the age of 14. On the other hand, a number of multi-day trips may partly be recorded twice: Once, when either the outward or the return trip is reported on the reporting day, twice, when the entire overnight journey is also reported in the 'travel module'. Accordingly, original microdata require several steps of post-processing, imputation and re-weighting in order to eliminate the effects of non-response

and potential double counts (or even over-reporting for example of air travel due to their perceived relevance). To this respect, a 'data fusion model' has been developed to generate a consistent trip dataset including trips and journeys of all distances for the entire German population, but without any overlap and double counts. Furthermore, the 'data fusion model' was calibrated on the basis of further socio-economic and official transport statistics.

The resulting trip file represents the core quantity structure to determine key figures for long-distance travel (e.g., travel demand and volume, trip purposes, modal split), also in relation to everyday mobility.

Further data sources are used to complement the picture, in particular in terms of data that are usually not provided by traditional NHTS (e.g., public transport passenger statistics). The compilation of data from different sources is usually a major challenge: Each data source is collected for a particular purpose, uses different conceptual definitions (e.g., single trip vs. complete journey including outbound and return trip) or threshold values to identify long-distance trips (e.g., a minimum of 50, 100 or 300 km one-way distance vs. a minimum number of overnight stays), and – last but not least – follows different methodological approaches (e.g., passenger counting vs. questionnaire-based trip interviews), which in turn may cause additional methodological bias.

So far, comparatively little information is available on background, motives and preferences of long-distance travelers, including information on modal choice or – as perceived from the traveler's subjective point of view – possible alternatives to actual travel decisions. Two ongoing projects on behalf of the German Federal Environment Agency seek to gain such additional insight by means of two dedicated surveys to complement the data framework described above: While the first one focuses on the use of particular transport modes typical for long-distance travel (e.g., camper vans, ferries), the second one aims to reveal insight in attitudes, preferences and travel-related decision-making.

The objective of this paper is fourfold: First, the challenge of data merging is described against the background of diverging definitions of long-distance travel. Secondly, the adopted approach of 'data fusion' is presented. Thirdly, the current picture of long-distance travel in Germany is outlined. Finally, methodological issues and challenges of data collection and data merging focusing on long-distance travel are discussed.

## **Keywords**

long-distance travel

data consistency

data fusion

data merging

## **Number of words**

*836 words (without title, authors, and keywords)*

## **Submission**

*June 15, 2019*