Underrepresentation of Work-related Trips in Traditional Household Travel Surveys

Anna Reiffer, Martin Kagerbauer, Peter Vortisch

Abstract

Travel behavior analyses and travel demand models still rely on data from household travel surveys which are often prone to underreporting of trips. Furthermore, traditional household travel surveys often do not allow for reporting of more complex work-related travel patterns due to limited trip purposes. To circumvent this issue, researches have chosen to conduct dedicated commercial travel surveys, however, surveys are often designed independently of existing data sources. And, while household travel surveys fail to provide the needed level of detail, they do account for some commercial activity. This paper clarifies which trips are covered in traditinoal household travel surveys and which travel patterns tend to be under-represented. The results of a cluster analysis conducted using both traditional household survey data and commercial travel survey data show that there are distinct work-related travel patterns. The analysis shows that not all travel patterns are represented in the traditional household travel survey: individuals that travel long distances on work-related trips or conduct a large number of work-related trips are scarcely accounted for.

1 Introduction

To this day, travel behavior analyses and travel demand models still rely on data from household travel surveys. Although information and communications technology (ICT) and especially GNSS technology have simplified the survey process in some cases, many traditional and nationwide travel surveys are still paper-pencil based. In these cases, the issue of underreporting trips weighs heavier as there are no mechanisms to validate trip characteristics, such as number of trips, trip start times, and trip distances. Previous research shows that work-related trips are affected by this underrepresentation [14, 15].

Furthermore, work-related trip purposes are often summarized within one purpose. However, commercial travel is just as complex as its private counterpart. For example, tradespeople tend to make trips with several different purposes: they may provide a service to a customer, transport material to a construction site or make a shopping trip to purchase material. While these different purposes entail different behavioral travel patterns, traditional household travel surveys only account for these trips using a single trip purpose [2, 7, 9, 16].

To circumvent this issue, researches have chosen to conduct dedicated commercial travel surveys [4, 18]. These surveys provide much more detail and allow for additional analyses regarding commercial travel. They also have proven to be suitable for commercial travel demand modelling [3, 5, 11, 12, 13]. While both commercial travel surveys and the travel demand models based upon them provide much needed information and insights into commercial travel behavior, the researchers developed their work independent of existing data sources (i.e. traditional household travel surveys) and travel demand models. However, while household travel surveys fail to provide the needed level of detail, they do account for some commercial activity.

This paper clarifies which trips are covered in traditinoal household travel surveys and which travel patterns tend to be under-represented. This study is structured as a comparison between data from a traditional household travel survey and data from a survey conducted specifically for commercial activities. Both surveys were conducted in Germany and within close temporal proximity. We applied cluster analysis to identify differences in the data and identify work-related travel patterns. We describe both the data and methods in the following section. The results section of the paper includes the outcome of our analyses, which we discuss in the subsequent section. The conclusion of this paper addresses the main outcomes of our work and implications for future work.

2 Materials and Methods

This study relies on two main sources of data: household travel survey data and commercial travel survey data. We describe those data and the processing below.

Data
To capture commercial travel patterns, the German Federal Ministry of Transport and Digital Infrastructure commissioned the nationwide vehicle-based travel survey Motorized Transport in Germany (*Kraftfahrzeugverkehr in Deutschland - KiD*). Recognizing that there are existing statistics and data sources concerning freight traffic produced by larger vehicles, KiD focused on commercial travel of light vehicles. The data comes in four different data sets: vehicle data, trip data, trip chain data, and geospatial data. KiD was carried out in 2002 and 2010. The sample of KiD 2010 - which we used as a databasis for our analyses of commercial trips - includes data on 70,249 vehicles, and on the survey day, 177,377 trips were conducted with these vehicles. [18]

As a proxy for traditional household travel surveys we used data from Mobility in Germany (*Mobilität in Deutschland - MiD*). MiD is a nationwide household travel survey commissioned by the German Federal Ministry of Transport and Digital Infrastructure with the purpose of capturing the households' daily travel behavior. Respondents were asked to report generic household information and their travel behaviour using a travel diary for one day. MiD has been conducted in 2002, 2008, and 2017. Although choosing the most recent data is generally sensible, we have opted to use MiD 2008 for two reasons: MiD 2008 is temporally closer to KiD 2010, which allows for a more stable comparison and MiD 2008 included a section on regular work-related trips. MiD 2008 includes four different data sets: car data, household data, person data, and trip data. For this study, we used the base sample, which includes information on 25,922 households, 60,713 persons, 193,290 trips, and 34,601 cars. [6]

Data preparation
While the data sets of MiD and KiD share characteristics, they are not comparable without adjustments. For our analysis, we first had to choose variables and then prepare and merge the data accordingly. As our goal is to identify travel patterns, we used trip related data. Regarding KiD, both the trip and the trip chain data set include trip related information. Although the trip chain data set includes fewer variables, they are already grouped by vehicle and summarize all relevant characteristics of the trip chains: cumulative travel time, cumulative activity duration, time of first work-related trip, cumulative trip distance, number of stops and trip purpose. While the data already comes in a prepared format, it includes a lot of missing values (see Table 1). Activity duration is the variable with the most prominent number of missing values. Although we recognize that this variable might have high explanatory value, we chose to exclude it from subsequent analyses, as we would lose too many observations. While cumulative travel time does not present as many missing values as activity duration, we opted for not using the variable either due to it's correlation with the variable cumulative trip distance. Subsequently, we removed entries with missing time of first trip and cumulative trip distance, resulting in 27,306 observations.

Table 1: Missing values in KiD trip chain data

Number of observations before cleaning data	27,677
Missing values by variable	
cumulative travel time	8,171
cumulative activity duration	25,377
time of first trip	130
cumulative trip distance	241
number of stops	0
Number of observations after cleaning data	27,306

According to the chosen variables in KiD, we selected the corresponding variables in MiD. For this step, we selected only work-related trips and determined the start of the first trip. We then grouped the data by respondent and summarized distances travelled and number of stops. This resulted in 1,157 observations. After this step, we were able to merge the data into a larger data set with 28,463 observations, which we used in the following analysis.

Analysis
To identify travel patterns regarding work-related trips and analyze differences between traditional household travel survey data and designated commercial survey data, we used cluster analysis.

Cluster analysis is a way to find groups in a population comprised of individuals by increasing the similarity within a group and the dissimilarity to other groups [17]. To account for the fact that the variables are measured in different units (time, distance, absolute numbers), we first need to scale the data. Otherwise cumulative trip distance, for example, would be weighted higher in the analysis than the time of the first trip as their ranges of value differ substantially. After scaling the data, we used the euclidean distance to calculate the distance matrix. This distance matrix serves as the input for the clustering analysis. There are many different clustering methods and the choice is highly dependent on the desired outcome. In our case, we did not want to make a priori assumptions about the number of clusters and we wanted the individuals grouped together to be as similar as possible. To maximize this similarity, the classical sum-of-squares criterion is optimal, as it minimizes within-group dispersion. The only agglomerative method that uses this criterion is *Ward's method*, which is the algorithm we chose for our approach. We conducted the cluster analysis in *R* using the *stats* package [10]. The package provides a function called *hclust* where the most common hierarchical clustering algorithms are implemented. As stated before, we applied Ward's method which is implemented in hclust as the method *ward.D2* [8]. The cluster analysis results in a dendrogramm which is a visual representation of the points at which the clusters are merged. This serves as a basis for determining the number of clusters, as it represents the distances between different cluster solutions. After determining the number of clusters, we analyzed the characteristics of each cluster and compared them to each other.

3 Results

Figure 1 shows the dendrogram of the cluster analysis. The graph shows that there are several possible solutions: Both 3- and 4-cluster solution are sensible as well as a 6- and 7-cluster solution. Due to the ambiguity, we first performed analyses of all cluster solutions and compared results. The comparison revealed that the 7-cluster solutions showed the most distinct results.

Figure 1: Dendrogram of ward cluster analysis

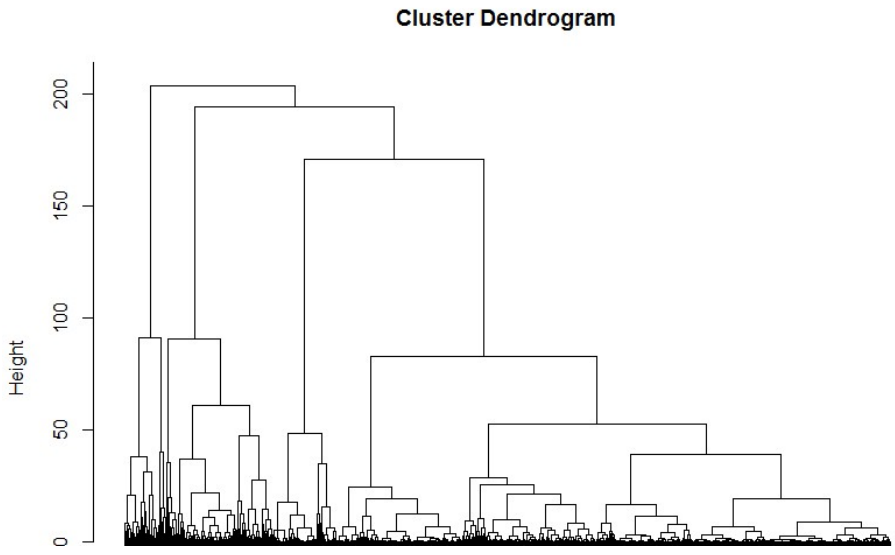

Table 2 presents the characteristics of each cluster of the 7-cluster solution. While all groups contain observations from KiD, both cluster 6 and cluster 7 do not contain any observations from MiD. The largest cluster is cluster 4, with a little over half observations falling into the group. It contains observations from both KiD and MiD. Regarding cumulative trip distance, cluster 3 and cluster 5 show much higher mean and median distances than other clusters. Observations in cluster 3 result in a mean distance of 366.01 km and those in Cluster 5 have a mean distance of 909.4 km, while all other clusters present mean cumulative distances between 39.62 and 110.52 km. Cluster 6 and cluster 7 are the most distinct groups considering total number of trips.

While all other clusters show an average number of trips below ten and a median number of trips below five, observations in cluster 6 result in an average number of trips of 84.88 and cluster 7 contains observations that result in 200.82 average number of trips.

Table 2: Cluster characteristics

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
N [-]	4692	2345	3560	15893	434	1279	260
N [%]	16	8	13	56	2	4	1
KiD [-]	4325	2027	3472	15519	424	1279	260
KiD [%]	16	7	13	57	2	5	1
MiD [-]	367	318	88	374	10	0	0
MiD [%]	32	27	8	32	1	0	0
Cumulative trip distance [km]							
mean	39.62	57.85	366.01	65.58	909.40	96.46	110.52
median	26	20	341	45	827.5	51	65
variance	39.82	106.87	146.97	61.74	257.25	105.76	116.74
minimum	0.1	0.1	4	0.1	640	3	3
maximum	240	1028	947	350	2500	629	729
Number of trips [-]							
mean	3.17	2.33	5.54	4.71	3.61	84.88	200.82
median	2	2	3	3	2	82	182.5
variance	2.42	2.98	6.40	5.38	7.45	27.80	55.93
minimum	1	1	1	1	1	27	123
maximum	22	50	52	56	85	165	462
Start of first trip [h]							
mean	9.85	15.59	5.82	6.94	5.61	7.83	7.23
median	10	15	6	7	6	8	7
variance	0.99	2.48	2.73	1.00	2.79	2.43	2.03
minimum	9	13	0	2	0	0	0
maximum	13	23	16	13	22	23	15

4 Discussion

Our results show that both data sets and therefore, both survey methods capture different travel patterns. Based on the cluster solution, we have identified four groups with distinct travel patterns: there are average mobile workers (cluster 4), late starters (clusters 1 and 2), long distance travellers (clusters 3 and 5), and highly active workers (clusters 6 and 7). The groups that include two clusters all have one cluster with moderate characteristics and one with more extreme characteristics.

The average mobile workers (cluster 4) are well represented in both surveys. They tend to start their day early in the morning, suggesting that they start their first trip from home and not from their business location. The mean cumulative trip distance of this cluster suggests that the workers stay within the area of their own operation. They same is true for the late starters (clusters 1 and 2), however, the start hour of the first trip suggests that members start their work day at the office or location of business. Especially cluster 2 shows late start times and while cumulative trip distances are not distinctly low, workers in that cluster on average only conduct 2.33 trips. The long distance clusters (clusters 3 and 5) each have very high average cumulative trip distances, and cluster 5 does not contain any observation under 640 km. To manage such long distances, they start their trips early in the day. Although they average more trips a day than the late starters, they do not make distinctly many trips. Opposed to those are the highly active workers (clusters 6 and 7): both clusters include workers that make many trips. Not only is the mean number of trips high - 84.88 and 200.82 respectively - but also the minimum numbers are considerably higher in these clusters: workers in cluster 6 make a minimum of 27 trips and workers in cluster 7 conduct a minimum of 123 trips. While both clusters show higher cumulative trip distances compared to late starters, they cannot be considered long distance drivers, especially considering the high number of trips, which suggest that each trip is relatively short considering distance.

We have found seven distinct clusters, however, only five had representatives of the traditional household

travel survey. This shows that not all patterns overlap. While especially clusters 1, 2, and 4 are well represented in MiD, more active workers are not represented in the traditional household travel survey. A relatively small share of MiD observations are included in long-distance clusters and none are represented in clusters characterized by large number of trips.

Although previous studies have drawn attention to the underrepresentation of work-related trips [1, 14, 15], they do not allow for an absoulute evaluation of the sample, i.e. if the sample is representative of the population in regards to their work-trip patterns. It is not suprising that traditional household travel surveys do not represent all work-related travel patterns, as the samples are not quoted by industry sector or specific jobs. However, as this study has shown, certain individuals are not accounted for in traditional household travel surveys at all.

Having identified this gap in traditional household travel surveys and therefore, in travel demand models based upon them, future research now needs to focus on evaluating which individuals of a population are affected by the underrepresentation.

5 Conclusion

This study examines the underrepresentation of work-related trips in traditional household travel surveys. The results of a cluster analysis conducted using both traditional household survey data and commercial travel survey data show that there are distinct work-related travel patterns. However, the analysis shows that not all travel patterns are represented in the traditional household travel survey: individuals that travel long distances on work-related trips or conduct a large number of work-related trips are scarcely accounted for. The results reveal that there not only exists the problem of non-reported trips, but also some individuals of the population are not represented at all.

These findings demonstrate the complexity of work-related travel behavior and that traditional household travel surveys are not fully equipped to account for all travel patterns. The results implicate that researchers and transport planners creating travel demand models need to increase attention to work-related travel behavior and acknowledge that - depending on the area of study - traditional household travel surveys may not provide a complete sample of the population.

The work presented here is specific to Germany as data sources allowed for a comprehensive analysis. While we may expect similar effects in other countries, further research needs to confirm this assumption.

Continued efforts are needed to gain insights into commercial travel behavior by increasing the data pool through surveys that complement traditional household travel surveys.

References

[1] Werner Brög and Gerhard Winter. Untersuchungen zum problem der "non-reported-trips" zum personenwirtschaftsverkehr bei haushaltsbefragungen.

[2] Adam Evans, John Cummings, Matthew Slocombe, and Francesca Corvaglia. National travel survey: England 2017.

[3] John Gliebe, Ofir Cohen, and John Douglas Hunt. Dynamic choice model of urban commercial activity patterns of vehicles and people. *Transportation Research Record: Journal of the Transportation Research Board*, 2003(1):17–26, 2007.

[4] J. Hunt, K. Stefan, and A. Brownlee. Establishment-based survey of urban commercial vehicle movements in alberta, canada: Survey design, implementation, and results. *Transportation Research Record: Journal of the Transportation Research Board*, 1957:75–83, 2006.

[5] J. D. Hunt and K. J. Stefan. Tour-based microsimulation of urban commercial movements. *Transportation Research Part B: Methodological*, 41(9):981–1013, 2007.

[6] infas Institut für angewandte Sozialwissenschaft GmbH and Deutsches Zentrum für Luft- und Raumfahrt e.V. Mobilität in deutschland 2008 - ergebnisbericht.

[7] KiM Netherlands Institute for Transport Policy Analysis. Mobility report 2016.

[8] Fionn Murtagh and Pierre Legendre. Ward's hierarchical agglomerative clustering method: Which algorithms implement ward's criterion? *Journal of Classification*, 31(3):274–295, 2014.

[9] Claudia Nobis and Tobias Kuhnimhof. Mobilität in deutschland - mid ergebnisbericht.

[10] R Core Team. R: A language and environment for statistical computing, 2019.

[11] Anna Reiffer, Michael Heilig, Martin Kagerbauer, and Peter Vortisch. Microscopic demand modeling of urban and regional commercial transport. *Procedia Computer Science*, 130:667–674, 2018.

[12] Anna Reiffer, Michael Heilig, Eckhard Szimba, Jan Klenner, Martin Kagerbauer, and Peter Vortisch. Combining macro- and microscopic approaches to model commercial transport demand in an urban area.

[13] K. Stefan, J. McMillan, and J. Hunt. Urban commercial vehicle movement model for calgary, alberta, canada. *Transportation Research Record: Journal of the Transportation Research Board*, 1921:1–10, 2005.

[14] Peter Stopher, Camden FitzGerald, and Min Xu. Assessing the accuracy of the sydney household travel survey with gps. *Transportation*, 34(6):723–741, 2007.

[15] Peter Stopher and Stephen Greaves. Missing and inaccurate information from travel surveys: Pilot results.

[16] Transport for London 2018. Travel in london: Report 11.

[17] Andrew R. Webb, Gavin Cawley, and Keith D. Copsey. *Statistical pattern recognition*. Wiley-Blackwell, Oxford, 3rd ed. edition, 2011.

[18] M. Wermuth, C. Neef, R. Wirth, I. Hanitz, H. Löhner, and H. Hautzinger. Mobilitätsstudie "kraftfahrzeugverkehr in deutschland 2010" (kid 2010) - schlussbericht.