Link, Heike et al.

Working Paper

Combining GPS tracking and surveys for a mode choice model: Processing data from a quasi-natural experiment in Germany

DIW Discussion Papers, No. 2047

Provided in Cooperation with:
German Institute for Economic Research (DIW Berlin)

Suggested Citation: Link, Heike et al. (2023) : Combining GPS tracking and surveys for a mode choice model: Processing data from a quasi-natural experiment in Germany, DIW Discussion Papers, No. 2047, Deutsches Institut für Wirtschaftsforschung (DIW), Berlin

This Version is available at:
https://hdl.handle.net/10419/273500

DIW BERLIN

2047

Discussion Papers

Deutsches Institut für Wirtschaftsforschung 2023

Combining GPS Tracking and Surveys for a Mode Choice Model: Processing Data from a Quasi-Natural Experiment in Germany

Heike Link, Dennis Gaus, Neil Murray, Maria Fernanda Guajardo Ortega, Flavien Gervois, Frederik von Waldow and Sofia Eigner

IMPRESSUM

DIW Berlin, 2023

DIW Berlin
German Institute for Economic Research
Mohrenstr. 58
10117 Berlin

Tel. +49 (30) 897 89-0
Fax +49 (30) 897 89-200
https://www.diw.de

ISSN electronic edition 1619-4535

Papers can be downloaded free of charge from the DIW Berlin website:
https://www.diw.de/discussionpapers

Discussion Papers of DIW Berlin are indexed in RePEc and SSRN:
https://ideas.repec.org/s/diw/diwwpp.html
https://www.ssrn.com/link/DIW-Berlin-German-Inst-Econ-Res.html

Combining GPS Tracking and Surveys for a Mode Choice Model: Processing Data from a quasi-natural Experiment in Germany

Heike Link [a,*], Dennis Gaus [a], Neil Murray [a], Maria Fernanda Guajardo Ortega [a], Flavien Gervois [a,b], Frederik von Waldow [a], Sofia Eigner [a]

July 2023

Abstract

This paper deals with the data generation process implemented for an analysis of the impact of the 9-Euro ticket on mode choice. We discuss the assumptions made and procedures used to process a raw dataset that is based on GPS traces of individuals' movements and on survey data into the choice-set for a discrete choice model. Several steps of cleaning and merging are described in order to a) obtain a reliable dataset; b) define available modal alternatives with attributes such as distance, duration, and costs; and c) impute the travel purpose for each movement to form. Our main contribution is to show that a systematic analysis of the sample obtained at different stages of data processing is important to make sure that the final sample is unbiased. Furthermore, we contribute by analysing the difference between observed travel time and travel time calculated by routing tools such as Google Maps. We show that the often-employed approach of estimating RP based choice models on the basis of observed travel times for the chosen mode of transport but calculated travel times for the non-chosen alternatives can introduce a structural bias into the sample.

Keywords: Data processing, travel behaviour, GPS traces, discrete choice models, revealed preferences

JEL classification: C55, C81, R41

[a]) DIW Berlin e.V., Mohrenstraße 58, 10117 Berlin, DE.

[b]) ENSAE Paris, 5 avenue Le Chatelier, 91120 Palaiseau, FR.

*) Corresponding author. E-mail address: hlink@diw.de

1. Introduction

In March 2022, the German Government agreed on a package of measures to offset the impacts of the Ukraine war on private households' cost of living. Amongst this package, two measures related to the cost of transport: the 9-Euro ticket allowed travel on all local and regional public transport from June to August 2022 at almost no cost, and a temporary decrease of fuel taxes to the minimum allowed by EU regulation lowered fuel prices by around 10% during the same period.[1]

The 9-Euro ticket was available for each of the three months of June, July, and August 2022. It granted nation-wide unlimited second-class access to busses, subways, trams, and regional trains throughout the respective month at a price of 9€ and could be bought at all venues selling public transport tickets (online, vending machines, bus drivers, etc.). Seasonal tickets with a validity of more than one month (e.g., yearly passes or student tickets) were automatically valid as a 9-Euro ticket, and owners of such tickets received a compensation of their originally paid price by the difference to 9€. The 9-Euro ticket thus introduced a strong negative shock to public transport prices in June 2022, followed by a respective positive shock three months later, when prices went back up to their initial level (and in some regions even higher due to price increases). It provides a unique quasi-natural experiment for transportation research as it uncovers travellers' behaviour before, during, and after the availability of the ticket. Advanced econometric approaches such as discrete choice models and causal inference methods allow to identify mode choice changes, price elasticities, and other responses to the policy intervention. For establishing the required database for such methods, we have used a combination of mobile-phone tracking information and survey data of the tracked individuals. Our approach is thus a revealed preferences (RP) study.

Disentangling the effects of the 9-Euro ticket from other influences, such as the fuel tax reduction mentioned above, the loosening of restrictions caused by the Covid-19 pandemic, and the summer holidays is a challenging task. The overlap of these events makes it difficult to define a past period for comparison and to identify an unaffected control group. Moreover, the public transport system underwent temporary interventions during June, July, and August 2022. In some regions, additional trains were ordered on short notice to increase supply and densify schedules. Nevertheless, especially tourist destinations encountered instances of overcrowded trains, with some reaching a capacity where further passengers were not permitted to board.

[1] See section 4.2.2 for a description of the development of fuel prices in period of interest.

Furthermore, infrastructure problems led to the closure of certain lines for extended periods, resulting in delayed or even disrupted connections.

This paper deals with the data collection and processing implemented for an analysis of the impact of the 9-Euro ticket on mode choice. It discusses the assumptions made and procedures used to obtain a reliable dataset and defines available modal alternatives with attributes such as distance, duration, and costs to form the choice-set for a discrete choice model. The motivation behind this paper is the lack of published studies on data generation and treatment behind modelling applications. Most papers on travel behaviour and respective models (e.g., mode choice, route choice, choice of departure time, etc.) discuss the process of generating the underlying dataset rather sparsely, usually for reasons of limited space in journal papers. Our main contribution is to show that a systematic analysis of the sample obtained at different stages of data processing is important to make sure that the final sample is unbiased. Furthermore, we contribute by analysing the difference between observed travel time and travel time calculated by routing tools such as Google Maps. This is an important issue since RP-based choice models often use observed travel times for the chosen mode of transport but calculated travel times for the non-chosen alternatives. Our analysis points out that such an approach can introduce a structural bias into the sample and highlights the need for further research on this subject.

The paper is organised as follows. Section 2 summarises the available methods for deriving individual trip data from GPS traces and contextualizes our approach within these methods. Section 3 focuses on the observed behaviour and provides a comprehensive review of underlying assumptions required for cleaning the tracking data and combining them with survey data. In addition, this section describes the methodology used for classifying and imputing the trip purposes. Section 4 provides a detailed description of the approach employed to identify the availability of modal alternatives and their corresponding attributes. Section 5 analyses the characteristics of the samples obtained at the different stages of the data processing, with a particular focus on potentially introduced bias. Finally, Chapter 6 concludes the paper.

2. Methodology for deriving personal trip data from GPS traces

The use of GPS-based traces to analyse travel behaviour has been an emerging field in transportation research since the mid-90es (see Shen and Stopher, 2014 for a review). Starting with GPS devices attached to cars and gradually moving to wearable GPS devices and smartphone solutions, GPS traces have been used to supplement or replace self-reported trip data (travel diaries) obtained from traditional methods such as paper-and-pencil interviews (PAPI),

computer assisted telephone interviews (CATI), and web-based travel surveys. Previous studies revealed several discrepancies between GPS traces and the traditionally collected travel diaries. For instance, considerable differences between self-reported and actual trip duration, have given rise to a variety of studies attempting to identify the magnitude of this phenomenon (Kelly et al., 2013, Peer et al., 2014, Spurr et al., 2015). Differences between self-reported and observed outcomes also occur with respect to other attributes of travel alternatives such as distance (see Hernandez and Witter, 2015 for perceived versus actual distance to metro stations and bus stops in Santiago) and trip costs (see Link, 2015 for estimated versus actual trip costs of motorists in two German cities).

Beside the opportunity to collect larger datasets and more detailed trip information, GPS traces shed light on these differences and overcome several shortcomings of the traditionally collected travel diaries. This includes under- or overreporting of trip frequencies (see Bricka et al., 2012) such as the lack of short (walking) trips which participants often neglect to report. GPS tracking enables comprehensive recording of all undertaken trips, provided that individuals have the necessary tracking app installed on their devices and carry them along. However, these advantages come at a cost: First, the amount of raw data to be processed requires automated treatment routines and extensive data cleaning due to malfunctions of tracking apps or insufficient signal coverage in certain areas. Secondly, it is essential to establish assumptions and thresholds, as well as calibrated models, to determine the transport mode and trip purpose from the collected data.

The processing of GPS-based data required to obtain a dataset with complete information for a mode choice model involves the following working steps:

- Error recognition and removal of invalid data
- Combination of single movements into uniquely defined trips
- Imputation of travel mode
- Imputation of travel purpose
- Identification of available modal alternatives for each trip
- Imputation of attributes for both the chosen and the non-chosen alternatives (e.g., duration, cost, access/egress times, transfers)[2]

The available approaches for these steps can broadly be classified into three major methods: rule-based (or criteria-based) approaches, probability models (mainly discrete choice) and

[2] Data generation for the non-chosen alternatives is necessary for a discrete choice model, but not for other methods such as causal inference approaches.

machine learning. While rule-based procedures are most common for error recognition, there is no dominating method for transport mode detection and trip purpose inference, and the accuracy achieved by these methods does not suggest any methodological preference (see Lei Gong et al., 2014 and Nguyen et al., 2020 for detailed reviews). Even more approaches are available for defining the choice set perceived by travellers and considered in their decision-making process. They comprise probabilistic methods (Calastri et al., 2019), captivity models (Gaudry and Dagenais, 1979), and thresholds set exogenously by the analyst for distance, access, egress, and further attributes. In the latter case, the thresholds are either used to define potentially available alternatives as part of the dataset, or they are included in the utility function via penalties (Martinez et al., 2009).

We have chosen a rule-based approach for most of our data treatment, which is supplemented by probability models. Due to the lack of information from self-reported travel diaries, a direct assessment of the validity of our approach in comparison to others is not possible. However, we use data from a small-scale accompanying survey for plausibility and consistency checks. In addition, we compare our final sample with data from the German Mobility Panel (MOP, 2022) and the German National Travel Survey (Mobilität in Deutschland; MiD, 2017), although figures from the latter refer to 2017.

3. Observed travel behaviour

3.1. Description of the data sources

For our study, we had access to data of the GIM Traces panel, obtained from the market research provider *GIM Gesellschaft für innovative Marktforschung*. This panel comprises data from individuals representatively drawn from the German population by factors such as age, gender, household size and region. All participants have agreed to install a geolocation tracking app on their smartphones. The continuously logged GPS location information was pre-processed into movements (i.e., continuous changes of location with a start point, an end point, and no stops in-between) by the Swiss market research company *intervista*. Furthermore, *intervista* adapted two models that were previously established in the Swiss context to fit our German application. The first one combines the recorded GPS data (speed, roads/routes taken, etc.) with geospatial information (e.g., location of railway and bus stations) to identify the mode of transport taken for the recorded movement, differentiating between walk, bicycle, car, local public transport (bus, tram, subway), train, and airplane. The second model identifies person-specific movement patterns to differentiate whether movements are for commuting (i.e., connecting home and

regular place of work or education of a person) or not. Consequently, a movement observation includes, besides movement- and person-specific identifiers, the start location and end location (as GPS coordinates), the start and end time, the used mode of transport, and the trip purpose (commuting/other). The pre-processed tracking dataset consists of 1.94 million movements conducted by a total of 4,891 individuals over five months between May (i.e., the month before the price intervention) and September (i.e., the month after the price intervention) 2022.

Additionally, an accompanying small-scale survey among the tracked panellists was conducted in three waves during June, July, and August 2022, respectively. The survey aimed at collecting information on socio-economic characteristics (age, gender, education, household income, household size, number of children), travel behaviour, and availability of transport modes (possession of a driver's license, car, bike, motorcycle, etc.; access to public transport in close proximity; possession of seasonal tickets including the 9-Euro ticket and discount cards for public transport). In the final survey wave, we also asked the participants for their public transport experience during the three months and their willingness to pay for a subsequent ticket. In total, 2,509 individuals responded to at least one of the three survey waves, providing data on the time-constant characteristics and month-specific information for at least one month.

3.2. Processing and cleaning of the dataset

By definition, the movement data cannot account for changes in the mode of transport or short stopovers: It registers the items of a multi-step route (e.g., cycling to the railway station, continuing by train, and walking to the final destination) as individual movements, distorting the start and end points of the entire route. Therefore, the data processing and cleaning[3] started with combining consecutive movements – defined as movements conducted by the same person, with the same purpose (commute versus non-commute, as defined by *intervista*), a maximum of 45 minutes between the end of one and the start of the next movement, and a maximum of 200m between the end point of one and the starting point of the next movement – into multi-step routes or trips. In the following, we will use the terms route and trip synonymously, with a route or trip consisting of at least one movement and representing the main unit of observation for our analysis.

For trips with more than one mode of transport, the main-mode concept (for a discussion see Varela et al., 2018) was applied: the mode used for the majority share of the total route (in

[3] Error recognition and removal of invalid data on the level of raw tracking data was performed by *intervista* and is not described here.

Table 1: Validity Requirements for Observed Routes

	Unit	Walk	Bicycle	Car	Train	Bus/Tram	Plane
Min air-line distance between start and end	km			0.2			
Min Distance – mode-specific	km	0.2	0.2	0.2	0.5	0.2	150
Max Distance – mode-specific	km	5	25	-	-	30	-
Min Speed Distance – mode-specific	km/h	2	10	10	10	10	100
Max Speed Distance – mode-specific	km/h	10	35	150	210	80	1000
Max Observed/Air-line Distance Ratio Distance – mode-specific	n. a.	2.5	3	4	4	5	2
Distance thresholds for non-frequent/new modes	Km	-	5 (e-scooter)	25 (Taxi/Uber)	-	-	-
Thresholds for removal of multi-modal routes	Km	More than 10km not travelled by main mode					
	%	More than 50% not travelled by main mode					

km) was considered as the mode of transport for the entire route (main mode), while all other parts were considered as access and egress steps. This introduces a distortion for multimodal trips with a close to equal distribution among several modes of transport. Since the construction of multi-modal alternatives for an RP choice set is out of scope due to the combinatoric character, multi-modality is not considered in our application (see below).

The combination of multi-movement routes was followed by an extensive cleaning including the removal of circular routes (i.e., routes with identical start and end locations), routes with a distance of only few meters (which might be caused by measurement errors or walking around the house), and lengthy bike and walking routes (which are probably leisure activities).[4] Furthermore, observations with a recorded mileage significantly shorter than the air-line distance between the start and end locations were removed, as were routes of the opposite case where the recorded distance was a multiple of the direct way. In addition, plausibility checks regarding the allocated mode of transport were performed based on a list of plausibility thresholds for distance, average speed, and the ratio between recorded distance and air-line distance (Table 1). Observations not in line with these requirements were omitted from the dataset.[5] Besides these mode-specific restrictions, Table 1 shows the minimum air-line distance between start and end points, applied to all routes as a general condition. Furthermore, routes with a

[4] In all three cases, it is infeasible to model a mode choice, as either the destination of the route is unknown (in case of circular routes), or the route does not have a destination.

[5] Combined routes are compared against a distance-weighted average of the restrictions applicable to their individual trips.

foreign start or end location were deleted, as the tracking data only cover movements in Germany. Due to the problems in constructing a multi-modal non-chosen alternative, only those combined routes that consisted of a clear main mode and access/egress steps remained in the dataset (see Table 1).

The remaining roughly 898,300 observations were characterised by a highly uneven distribution across individuals. Individual participants have registered up to 1,356 routes, while 1,026 persons were tracked on a total of 30 trips or less during the entire 5 months (possibly due to leaving the house without their smartphone, deactivating the tracking feature, or temporarily uninstalling the tracking app).

Further cleaning was necessary for the survey due to minor inconsistent answers such as participants claiming to have worked more than 7 days during one week. Inconsistent combinations of answers regarding the possession of seasonal public transport tickets (e.g., the possession of a yearly ticket in July, but neither in June nor in August) were corrected based on assumptions, and cases remaining unclear were dropped. Finally, for the envisaged modelling of the effects of the 9-Euro ticket over the entire period, we established a balanced survey panel of 1,233 participants who provided consistent answers in all three survey waves.

3.3. Combination of tracking and survey data

The combination of tracking and survey data provides a wealth of information on people's travel behaviour that is unachievable with either one of the data sources alone. Additionally, the two sources can be used to crosscheck the validity and consistency of each other. However, combining these data also means that only the intersection of both sources can be used, reducing our dataset to 289,200 recorded routes conducted by 1197 participants that completed all survey waves. To ensure a consistent and comprehensive representation of the movement behaviour of these participants, which also includes days without any routes, the sample was further restricted to a set of continuously tracked participants: It includes 864 participants for whom a geolocation signal (i.e., not necessarily a route or movement) was received on at least 15 days during each of the five months. The validity check of the recorded routes referred to two issues: First, the plausibility of the travel mode assigned, and second, the treatment of "unusual" modes of transport, such as taxis, e-scooters, and Uber. To start with the first issue, we found that the assignment of travel modes through a probability model based on tracking information was plausible in the vast majority of cases. 7,900 inconsistent observations, mostly recorded in dense areas with multiple modes of transport using the same infrastructure and moving with

similar speeds (e.g., cars, busses, and bikes in busy cities) were omitted from the dataset since individuals reported to not have the respective transport mode available or never use it.

Identifying "unusual" modes of transport from GPS traces is in general a challenging task due to their identical movement patterns compared to traditional alternatives such as cars or bicycles: A car trip of a regular user of both cars and taxis cannot be clearly identified as car or taxi, and comparable difficulties arise with respect to bicycles versus e-scooters. Allocating these movements to traditional modes of transport, however, would introduce a bias into a mode choice models since they are characterised by significant differences in costs compared to traditional alternatives.[6] Therefore, for those respondents who reported regular use of both "unusual" modes and the traditional alternative, observations with an unclear mode choice were removed (see Table 1 for the assumptions on the average speed and maximum route length of uncommon alternatives). The dataset after cleaning and merging GPS traces and survey data comprises a total of 230,500 routes for which both the chosen and the non-chosen alternatives were calculated.

3.4. Determination of trip purpose

Trip purpose plays an important role in mode choice models, as the consequences of a price intervention vary depending on the purpose of routes. In the absence of self-reported purposes, it was neither possible to implement machine-learning techniques (which require a training sample) nor to assess the accuracy of the resulting route classification. Therefore, we used the distinction between work/education trips and others as determined by *intervista* through a probabilistic model as a starting point and employed a rule-based approach to obtain a more detailed definition of purposes. For an assessment of plausibility, we compare the results from our approach with the structure of (self-reported) trip purposes in the German National Travel Survey (MiD 2017). A central issue for imputing the trip purpose of a given route is the typology and the level of detail used to define different purposes (see Nguyen et al., 2020 for a summary of available types of purpose classifications). As a general requirement, the classification chosen should allow to fill each category with a relatively high degree of confidence based on the variables at our disposal. Furthermore, it should be sufficiently detailed to separately estimate the impact of price variations on various types of mobility. For purposes of comparability, our classification, which consists of nine categories and several subclasses (Table 2), aimed at being close to the one used in the MiD (2017).

[6] This difference would in fact require the introduction of respective modal alternatives in a mode choice model.

Table 2: Trip purpose categories

Category	Sub-categories
To home	
Commute	Commute education, Commute other
Transfer	
Professional	
Shopping	Daily shopping, Long-term shopping, Services, Fuel, Other
Private business	Medical, Other (bank, administration, car dealership, etc.)
Pick-up / Drop-off	Children to school/kindergarten
Leisure	Eating, Entertainment, Worship, Culture and events, Tourism, Sport, Nature, Allotment, Other
Visits	

We started by imputing a user-specific place of residence as the most visited geographical point and creating dummy variables indicating whether a route starts or ends at home. The categorisation obtained from *intervista* allowed us to identify commute routes. The remaining categories were filled using three data sources: route characteristics, user socio-economic characteristics, and geographic information obtained from the open geographic database OpenStreetMap (OSM). These variables include the employment status and the number of children (to identify professional and pick-up/drop-off trips, respectively) as well as points of interest, land use, and available transport infrastructure at the endpoint of the route. The points of interest (PoIs) comprise a broad range of amenities, such as shops, restaurants, museums, sport pitches, parks, places of worship, schools, and hotels. These were identified as potential destinations of a trip if the maximum distance from the endpoint was below 100 meters.

4. Determination and calculation of modal alternatives

A necessity in the process of generating an appropriate database for a (RP-based) discrete choice model is completing the choice set by identifying the non-chosen (available) alternatives for each observed route and calculating attributes for the chosen and non-chosen alternatives. These attributes commonly include trip cost, trip duration, and additional level-of-service attributes such as the access/egress and waiting times, the number of transfers, and the frequency of connections. In our case, the choice set consists of five modal alternatives: walk, bicycle, car, public transport, and flight. The availability of each mode for each route depends on the individual and the specific trip. To prepare the processed dataset for model estimation, the available modes for each route were determined and the corresponding attributes were calculated. The attributes of the observed trips were compared with the respective calculated

Table 3: Requirements for defining non-chosen alternatives as part of the choice-set

Mode	Access condition	Distance thresholds (air-line)
Walk		Max. 5 km
Cycling	Own bicycle or bike-share	Max. 25 km
Car	Own car available	No threshold
	Car sharing available	Max. 50km
Public transport	Access to public transport available	No threshold
Plane	Direct flights available between accessible pair of airports (max distance between airport and start/end location 100km)	Min. 150km

information to ensure that the calculated alternatives match the observed data. Following Tsoleridis et al. (2022) and Calastri et al. (2018), we focus on the calculated attributes (travel times, access/egress and number of transfers) for both the chosen and the non-chosen alternatives to ensure that the values in the choice set originate from the same data generation process.

4.1. Identification of choice set

The identification of non-chosen alternatives as part of the choice set was based on the observed distance of the chosen route and the information on mode availability from the survey. Table 3 summarises the distance thresholds and access conditions that a non-chosen alternative had to meet in order to be considered in the choice-set. For public transport, the origins and destinations of all observed routes were complemented by the closest passenger railway station (measured as air-line distance), with station information obtained from the station register of the Deutsche Bahn AG.

The availability of a flight option is determined by the proximity of airports to the origin and the destination as well as the existence of regular direct flights between these two airports. To achieve this, a list of potential airports is established based on all airports utilized for at least one flight connection in the tracked routes.[7] All start and end locations of recorded trips with no more than 100km (air-line distance) from airports in this list were considered accessible from the respective point. Using flight schedules published by the airports for winter 2022/23, direct flight connections between all airports were identified.

As a general note of caution, we have to mention that the availability of a mode as perceived by the individual might differ from mode availability as defined by our thresholds (see for example Schmid et al., 2022 who state that RP data often tend to over-estimate availability

[7] These were the following airports: Berlin, Bremen, Dresden, Düsseldorf, Frankfurt, Halle/Leipzig, Hamburg, Hannover, Köln/Bonn, München, Saarbrücken, Stuttgart.

of certain modes). With our approach, we account for the individuals' general availability of travel modes obtained from the survey. However, individuals might perceive mode availability depending on attributes such as time and cost and not only availability or proximity of modes, which we cannot account for in absence of any information on trip-specific circumstances.

4.2. Calculation of choice set data

We combine information from various sources such as OSM and OpenRouteService (ORS), Google Maps, the Deutsche Bahn website, and Skyscanner to calculate the distance, duration, price, and further level-of-service variables such as access/egress times and the number of transfers for all chosen and non-chosen transport modes.

4.2.1. Travel distance and duration

The main source of information for walk, bike and car routes was a local ORS installation using OSM data for Germany from September 1st, 2022. This routing tool calculates connections between any given pair of coordinates while accounting for potential restrictions (e.g., speed or access restrictions based on type of road) and provides the travel distance and travel time of the fastest connection.[8] While this approach produced reliable results for foot and bike routes, it should be noted that for car routes, the ORS employs an algorithm that does not account for traffic interactions or congestion, providing travel times based on the optimal case (cf. Tsolerides et al., 2021).

Distance and duration of public transport were derived from Google Maps via the Maps API. Assuming that the start time of most trips is flexible within a window of up to two hours, public transport connections between the start and end locations at the observed weekday and time were requested from Google Maps and distances and travel times of all available connections within the defined time window were averaged.[9]

The distance of flight options was approximated by the air-line distance between the start and end locations. The duration was determined through a more complex procedure, combining the average flight time obtained from Skyscanner via an API with an additional hour for procedures at the airport (check-in, security check, boarding) and an access and egress time. To quantify the access and egress times, the distances between the start location and the departure

[8] The ORS settings were retained from the default values with the exception of higher maximum route lengths.
[9] As Google Maps cannot provide connections in the past, the same weekday three weeks from the day of calculation was used. This might introduce a slight distortion due to changed schedules in individual cases.

airport as well as between the arrival airport and the final destination were assumed to be covered with an average of 40km (air-line distance) per hour.

4.2.2. Costs

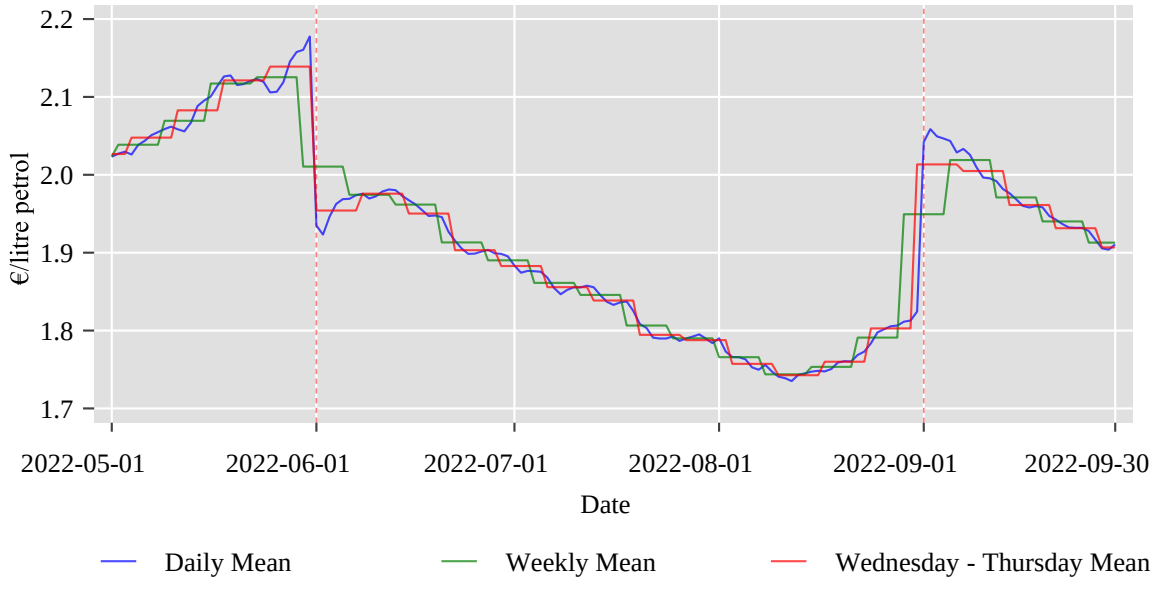
a) Walking and cycling

Walking as well as cycling with an own bicycle was assumed to have zero costs. Individuals without an own bicycle but with a bike-sharing station in their proximity were assigned a cost of 0.10€ per minute of travel time for bike options.[10]

b) Car

In contrast to parts of the available literature (Tsoleridis et al., 2022; Varela et al., 2018), we considered only fuel costs for car trips, justified with the argument that short-term travel decisions are based on perceived (out-of-pocket) costs rather than costs of vehicle purchase, maintenance, insurance, etc. Furthermore, the sheer variety of parking rules and fees prevented the inclusion of parking costs in a nation-wide study like ours. As data protection measures in Germany do not allow a direct link between survey participants and the national vehicle registry (see La Paix et al., 2021 for such an application in the Netherlands), we do not have information on the specific type of car used by the individuals. Consequently, it is not possible to infer the route-specific fuel consumption and costs. As an approximation, we calculated trip-specific fuel costs based on three data sources.

Figure 1: Average gasoline prices per day, calendar week and week from Wednesday to Thursday

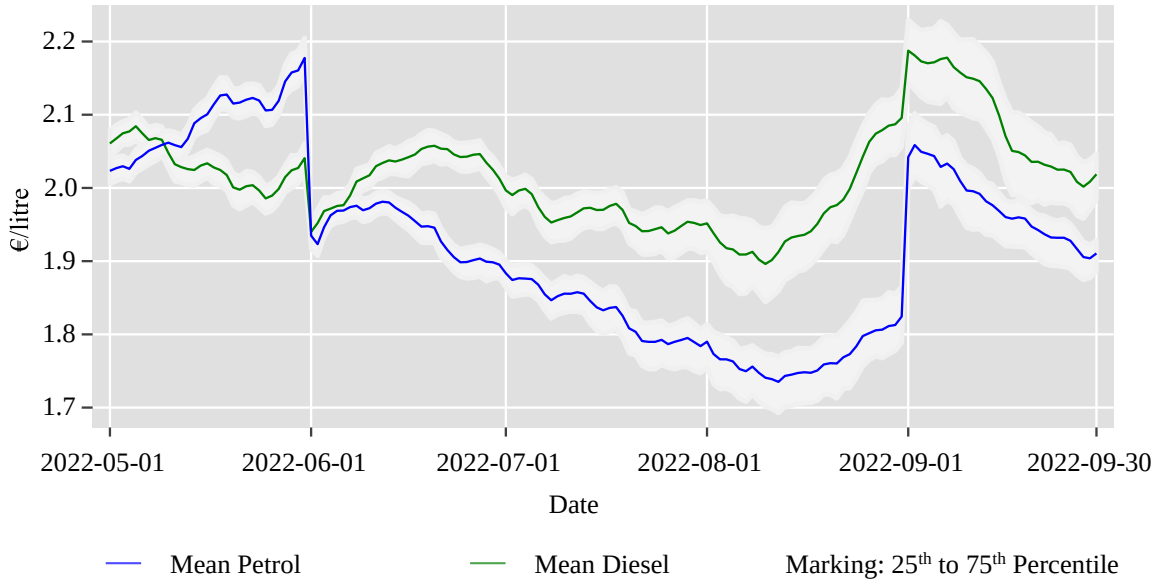


[10] The two largest bike sharing companies in Germany raise fees of 1€ per 15 minutes (https://www.nextbike.de/de/preise/, https://www.callabike.de/de/start/)

First, we used fuel price data from the "Tankerkönig" API, describing prices of petrol and Diesel at all petrol stations in Germany during the entire period.[11] While it is unclear whether daily or weekly prices affect individual's fuelling and travel behaviour, Figure 1 shows a clear effect of the fuel tax reduction on Wednesday, 1 June 2022, and on Thursday, 1 September 2022, supporting the use of daily prices. Relying on weekly prices may lead to underestimating anticipation effects of the fuel discount. Figure 2 shows the considerable spatial variation of fuel prices in Germany, suggesting the use of disaggregated data. Furthermore, petrol stations at motorways charge higher prices than other stations with a difference of around 0.25€/litre (see MTS-K, 2021), but play a minor role for the daily fuelling behaviour of individuals. We therefore eliminated these stations from the dataset based on address and geolocation data to avoid price distortions. The remaining data obtained from Tankerkönig were aggregated to county-specific (NUTS-3 level) daily average prices of gasoline and Diesel.

Second, information on the average fuel consumption of different types of vehicles was collected from the MOP (2022) from 2019 to 2022.[12] Since the MOP determines fuel consumption through individual fuelling diaries, consumption values align with real conditions. Third, these data were combined with regionally differentiated (NUTS-1 level) information on the fleet composition obtained from the National Vehicle Registry (KBA, 2022). The fleet compo-

Figure 2: Daily average prices for Diesel and gasoline (E5 and E10) in all NUTS3-Regions



[11] "Tankerkönig" is an information service for consumers listed at the Market Transparency Unit for fuels (see https://www.bundeskartellamt.de/EN/Economicsectors/MineralOil/MTU-Fuels/mtufuels_node.html). Petrol station operators and oil companies are obliged to provide real-time data on fuel prices per station. The data used in this paper were obtained under the Creative-Commons-License (CC BY 4.0).

[12] MOP (2022) fuel consumption figures were kindly provided by the Karlsruher Institute for Technology (see https://daten.clearingstelle-verkehr.de/192/)

sition corresponds to the vehicle classification used in the fuel consumption data and differentiates between types of drives. Connecting these two data sources thus allows us to derive a state-specific "average car" as a weighted average of vehicle classes and types of drives. By combining this information with the strongly disaggregated fuel price data, day- and county-specific average costs per km were calculated.

For the few individuals who do not have access to a private car, but to a car sharing alternative, car costs were approximated with 0.30€ per minute for trips with a distance of up to 50km.[13]

c) Public transport

Public transport prices were collected from the website of Deutsche Bahn using web scraping, considering person- and trip-specific factors. Generally, it must be noted that the tariff structure of public transport and especially train travel in Germany is extremely complex. On the one hand, the local and regional public transport is managed by almost 80 transport associations (so called "Verkehrsverbünde"), public authorities with a coverage ranging from entire states to single counties. Each area has its own prices, tariff structure, seasonal tickets, special offers, and cooperation between some regions allows for special prices when travelling into neighbouring areas, while crossing other borders is covered only by a relatively expensive federal tariff. In addition, there are flat-rate tickets allowing unlimited travel for an individual or a group within a certain area (e.g., a state) and time (e.g., one day) for some regions. On the other hand, long-distance train traffic is almost exclusively run by Deutsche Bahn, applying a rather non-transparent pricing structure. For all connections in this segment, tickets are available at a flexible standard price depending on the types of trains used, weekday and time, and other factors. In addition, different types of discounted tickets exist, depending, besides others, on the specific connection and time of booking. Furthermore, different types of Bahncards (discount cards) grant a discount of 25%, 50%, or 100% on all long-distance train tickets, but are accepted only in few of the transport associations.

This intricate pricing structure has two significant drawbacks: First, a trip-specific price should be used whenever possible, posing a significant computational cost, as it is difficult to approximate prices with an algorithm (e.g., a price per km). Second, the exact price an individual paid (or would have paid) for a specific trip cannot be identified retrospectively. To address this issue, we utilized web scraping to gather all available prices of three public transport

[13] For longer trips, car sharing was not considered as an available option. For a comparison of carsharing prices, see https://www.bussgeldkatalog.org/carsharing-kosten/#was_kostet_carsharing_ein_preisvergleich.

Table 4: Model estimation for public transport prices

	Local/Regional	Long-Distance
Intercept	1.89***	11.51***
	(0.03)	(0.10)
Distance	0.28***	0.13***
	(0.00)	(0.01)
$Distance^2$	-0.00***	-0.00***
	(0.00)	(0.00)
Within_county	-0.19***	
	(0.03)	
R2	0.80	0.47
Observations	55,223	1,552

Notes: Standard errors in parentheses.- *** marks significance at the 99% confidence level.

connections comparable to the observed routes. Comparable connections were defined as those departing a) on the same weekday three weeks into the future from the request, b) no more than one hour earlier or later than the observed movement, and c) connecting the start and end stops identified by Google Maps. Despite implementing multiple checks, this procedure occasionally results in routes that differ from the connections provided by Google Maps (e.g., because a bus station name occurs in multiple cities). To minimize the chance of such mismatches as well as to reduce computation times, we restricted the requests to connections between railway stations (instead of bus or trams stops) for longer routes.[14] For each comparable connection found, the lowest available price was chosen, and the final price was calculated as the mean price across the connections. The obtained 55,000 prices for local and regional connections and 1600 prices for long-distance train journeys were used to estimate two quadratic models of a price per kilometre (one for local and regional transport, one for long-distance connections) to impute costs for around 34,000 connections retrieved from Google Maps for which Deutsche Bahn did not provide a price. The estimation results of these models are summarized in Table 4.

With respect to seasonal and discount cards, we used the same approach as for car costs by considering only the marginal costs of each trip for the individual. This means that the purchase cost of seasonal tickets or discount cards was not considered in the cost attribute for public transport, while discounts granted by those cards were incorporated into the price where applicable. Bahncard discounts were only considered for routes including long-distance trains. For individuals with a seasonal ticket, a zero price for trips within the same region was considered. [15] Trips conducted during June, July, or August by local and regional public transport only

[14] In this case, access and egress costs (0.45€ per air-line km) and times (with a speed of 40 air-line km per hour) were added to the connection using the same method as applied to flight connections.
[15] Based on a subsample analysis, it was assumed that routes with an air-line distance of up to 15km are within one tariff region and thus covered by local seasonal tickets.

were allocated a maximum price of 9€ if the individual did not own a 9-Euro ticket, and a zero price otherwise.

d) Airplane

The logic of comparable connections (see section c) was also used to determine a price for flight alternatives, with prices requested via a Skyscanner API. For each comparable connection, defined as direct flights between a pair of reachable airports (see section 4.1) on the same weekday three weeks from the request, the obtained lowest available price was averaged. Costs and time for access and egress were added assuming a flat cost of 0.45€ per air-line distance kilometre and a speed of 40 air-line km per hour.

4.2.3. Further level-of-service variables

Passenger's experience of a trip – especially in public transport – depends strongly on the so-called "level of service", which describes for example access and egress steps as well as waiting times and transfers. As this concept refers particularly to the public transport system, we fixed the respective values to 0 for walking, cycling, and car. For public transport trips, we extracted further trip characteristics from the Google Maps API. First, we obtained the number of transfers of a route, independent of the time and distance they consist of.[16] Second, the data allowed us to distinguish between in-vehicle time (IVT) and out-of-vehicle time (OVT) for public transport, which for our route data account on average for 56% and 44% of the total trip duration, respectively. Additionally, we disaggregated the OVT into access, egress, transfer, and waiting times. The access time refers to the walking time to reach the starting public transport stop of a route and comprises nearly 43% of the OVT. Egress time, describing the walking time

Table 5: Validity Requirements: Comparison Observed/Calculated Routes for recorded trips

	Unit	Foot	Bicycle	Car	Public Transport
Min Ratio – Distance[1]	-	0.67	0.5	0.57	0.5
Max Ratio – Distance[1]	-	1.5	2.0	1.75	2.0
Min Ratio – Duration[2]	-	0.67	0.5	0.67	0.4
Max Ratio – Duration[2]	-	1.5	2.0	8.0	2.5
Min Distance Difference	km	0.5	0.5	1.0	2.0
Min Duration Difference	Min	10	10	10	10

Notes: [1] Ratio calculated as observed/calculated distance.- [2] Ratio calculated as observed/calculated duration.- Routes are invalid if distance ratio and distance difference are outside limits or if duration ratio and duration difference are outside limits.

[16] This definition means that walking from one station after getting off a vehicle to another station before boarding another vehicle counts as one transfer.

from the last transit service to the destination, is slightly shorter on average than access time, representing 40% of the OVT. We considered transfer time as the walking time for interchanges between two transit vehicles that is necessary to get from one vehicle to the next. Consequently, transfer time does not include waiting time, which explains why it represents only a small percentage of the OVT (2%). Finally, waiting time for transit vehicles represents on average 15% of the OVT.

Since domestic flight distances in Germany are generally so short that in many cases the access to and egress from the airport takes longer than the flight itself, only direct flights were considered as a choice option.[17] While this implies zero transfers between flights, access to the departure airport and egress from the arrival airport are considered as one transfer each, assuming that they consist of only one step each.

4.2.4. Further control variables

To account for the impact of weather conditions on mode choice, day- and county-specific weather information was assigned to each trip. For this purpose, we used data obtained from the Climate Data Centre operated by the German Weather Service.[18] The measurements of 5558 temperature stations and 493 precipitation stations, captured continuously between May and September 2022, are geolocalised and were aggregated into daily averages at the NUTS-3 level to correspond to our sample of routes.

4.3. Comparison between calculated and recorded data

As a final step in plausibility checking and cleaning, we compared the distance and duration of the chosen alternative (observed routes) with their calculated counterparts. Routes where the observed trip attributes differed strongly from the calculated values – either because the individual did not move directly from the start to the destination or due to a wrong mapping in the calculation – were removed. Table 5 shows the assumed thresholds per mode as minimum and maximum ratio between observed and calculated distance and travel time. Furthermore, a maximum absolute deviation was defined for these trip characteristics. Since especially car routes can be subject to a significant difference between calculated and observed travel time due to delays and congestion (see Section 5), the validity threshold for travel time by car was set

[17] As the longest possible flight is about 1:15 hours, an additional transfer between flights would increase total travel time unjustifiably.

[18] The Climate Data Center (CDC) offers extensive weather data from local measuring stations (see https://cdc.dwd.de/portal/)

considerably higher than for other modes.[19] By dropping routes violating the threshold, we obtain a final sample of 203,500 routes conducted by 864 individuals.

5. Findings

5.1. Characteristics of the sample at different stages of data processing

While our initial dataset included 4811 panellists conducting 898,300 routes, the final data used for the mode choice modelling consists of 864 individuals and 203,500 routes. This reduction, caused by the necessary data cleaning, the combination of data sources, and the requirement of a balanced and consistently tracked panel over the period from May to September (i.e., before, during and after the policy intervention), is a common experience when working with tracking data (see for example Bansal et al., 2021; Tsoleridis et al., 2022).[20] In this chapter, we analyse whether potential shifts in the level and/or structure of variables have introduced a bias into the final sample.

In terms of trip frequency, the number of average trips has slightly increased throughout the process, reaching 235 routes per person and 1.54 trips per person and day in the final sample (Table 6). The average trip length and duration as well as the distribution of trip purposes remained almost constant. It must be noted that the MiD (2017), where respondents report their daily trips in large detail in a one-week travel diary, shows more trips per person and day as well as longer trip distances than our sample. This holds also for the data from the German Mobility Panel, referring to 2021, which gives higher figures for the number of trips, trip length and trip duration. An explanation of these differences has to consider several aspects: First, our definition of the term "trip" differs from those in the MiD (2017) and the MOP (2022) due to our data generation process combining single movements to a trip.[21] Second, both the MiD (2017) and the MOP (2022) are self-reported surveys with presumably more weight given to longer trips by respondents than in our data. Third, the MiD (2017) refers to a pre-Covid period with more trips than in our data from 2022. The model split was hardly impacted by the cleaning process: The shares of all modes of transport remained almost constant and are very close to the ones observed in the MiD (2017). The same holds for the availability of travel modes, while the share of participants with a seasonal ticket in May as well as with a 9-Euro ticket in the other months decreased slightly.

[19] However, it should be noted that delays and congestion also affect travel times of public transport.
[20] Tsoleridis et al. (2022) lost one quarter of individuals and almost 80% of trips in processing their cross-sectional data. Bansal et al. (2021) even suffered a loss of 85% of individuals in their study.
[21] If route legs are not combined to routes, the number of trips per day and person in our sample amounts to 2.6.

Table 6: Characteristics of trip sample at different stages of data processing

	Cleaned routes[1]	Unbal-anced panel[2]	Balanced panel[3]	Final sample[4]	MiD (2017)	MOP (2022)
Individuals	4811	2429	1197	864	316,000 [a)]	3247
Number of routes	898,288	525,913	289,239	203,450	-	-
Number of trips per day and person	1.22	1.42	1.58	1.54	3.1	2.94
Trip characteristics						
Average trip duration (min)	25.7	25.8	25.6	24.8	-	75.0
Average trip length (km)	12.6	12.7	12.7	13.3	39.1	35.9
Trip purpose (%)						
Work/education	20.6	20.6	19.4	19.9	23.4	-
Others	79.4	79.4	80.6	80.1	-	-
Modal split of routes (%)						
Walk	22.4	22.7	22.7	22.5	21.6	
Cycling	12.1	11.9	13.2	11.7	11.0	
Car	53.7	53.6	52.1	56.1	57.1	
Public transport	11.8	11.8	12.0	9.7	10.2	
Air[5]	0.0	0.0	0.0	0.0	0.0	
Mode availability						
Car (%)		88.3	88.9	88.5		
Public transport (%)		85.0	85.2	85.5		
Ticket availability (%)						
Seasonal ticket - May		28.1	25.8	25.6		
9-Euro ticket - June			46.7	44.9		
9-Euro ticket - July			47.6	46.1		
9-Euro ticket - August			46.0	44.2		

Notes: [1] All tracked individuals incl. those not having responded to the survey.- [2] Individuals who participated in at least one survey wave and have at least one tracked route.- [3] Individuals who participated in all 3 survey waves and have at least one tracked route.- [4] Individuals who participated in all 3 survey waves, were tracked consistently (at least 15 days/month) between May and September, and have at least one recorded route.- [5] Mode share of aviation was below 1%.- [a)] Final sample. Individuals correspond to 156,000 households surveyed.

With respect to socio-economic characteristics (Table 7), our data processing has led to an increase of the mean age at the expense of individuals below 35 years. In addition, male respondents are slightly over-represented. This is confirmed by a comparison with data from the German Socio-economic Panel (SOEP, 2022), which is a fully randomly drawn sample without any quota design, whereas our original panel was representatively drawn based on quotas. However, the data processing did not lead to major shifts with respect to income distribution and occupational status: In general, we observe the income distribution to be highly similar to the SOEP data, with a slightly lower non-response rate. However, the categories for the occupational status are not fully comparable, as the SOEP-category “at home/not occupied” contains

Table 7: Socio-economic characteristics of sample at different stages of data processing

	Unbalanced panel[1]	Balanced panel[2]	Final sample[3]	SOEP (2022)[4]
Individuals	2509	1233	864	32,022
Age				
Mean	44.3	47.9	47.7	44.0
Categories (%)				
below 18	0.5	0.3	0.2	4.3
18 – 35	29.6	19.8	19.2	33.6
36 – 60	55.2	60.0	62.5	43.2
above 60	14.7	20.4	18.1	21.2
Sex (%)				
Male	51.3	55.2	56.6	50.6
Female	48.7	44.8	43.3	49.4
Income groups (%)				
Not declared	5.4	4.5	4.4	7.0
below 1000€	5.1	5.4	4.6	5.2
1000-2000€	18.9	18.7	17.1	18.0
2000-3000€	24.4	24.6	23.4	21.4
3000-4000€	21.8	22.1	23.3	17.5
above 4000€	24.4	24.8	27.2	30.9
Occupation (%)				
Full-time worker	58.8	55.2	58.9	38.6
Part-time worker	16.3	15.7	15.2	14.4
Education[5)]	3.2	3.7	3.0	3.4
Retired	13.2	17.1	15.3	13.9
At home/not occupied	8.8	8.2	7.6	38.7
Household size				
Number of persons in household (mean)	2.48	2.35	2.32	3.13
Children (<6 years) (%)	15.6	13.3	13.4	15.0
Children (at school) (%)	24.4	22.6	21.3	33.8

Notes: [1] Individuals who participated in at least one survey wave; includes 80 individuals without observed routes.- [2] Individuals who participated in all 3 survey waves; includes 36 individuals without observed routes.- [3] Individuals who participated in all 3 survey waves, were tracked consistently (at least 15 days/month) between May and September, and have at least one recorded route.- [4] Sample consists of individuals older than 15 years.- [5] School, university, apprenticeship, job training.

also retired individuals and the category "Education" in the context of occupation refers only to apprenticeships. Bearing this in mind, our sample has a larger share of full-time workers than the SOEP at the expense of unemployed persons, individuals at home, and retirees. Furthermore, households in the SOEP are larger than in our sample, and a larger share of them has children. While the difference is small for the share of households with children below six years

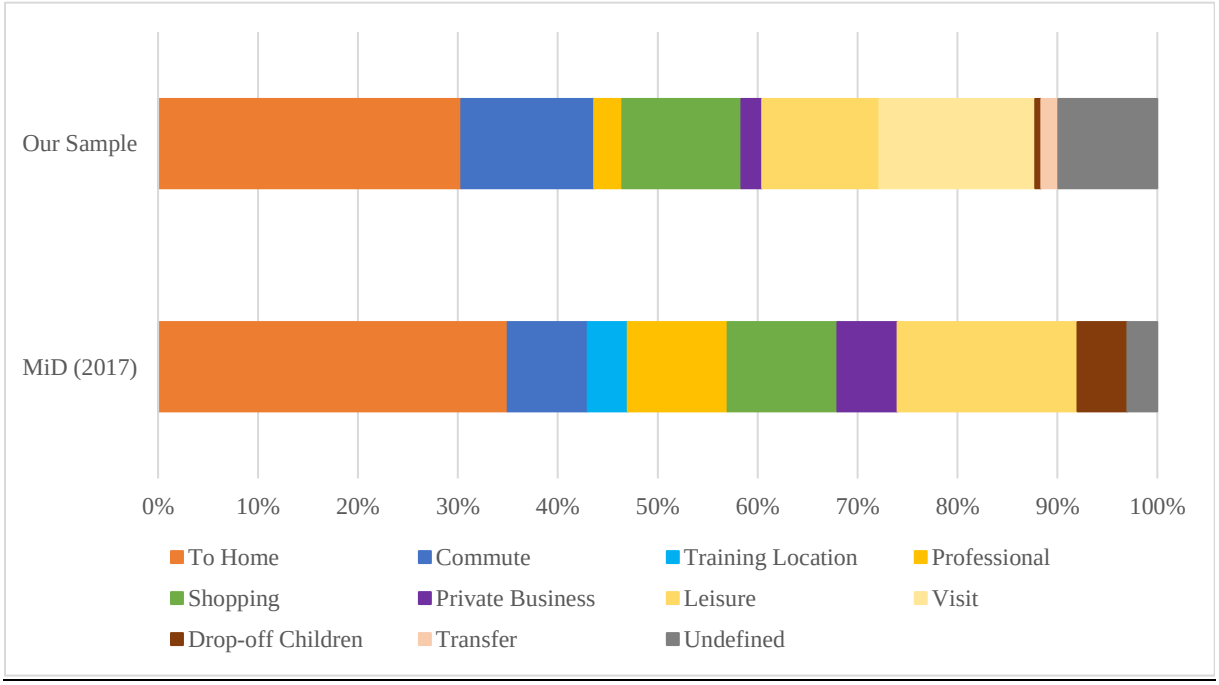
of age, our sample underrepresents households with children at school. However, this is also the case for our unbalanced sample and has not been introduced through the data processing.

Overall, it can be concluded that the process of combing movements into routes, cleaning the data, and restricting the sample to individuals having responded to all survey waves and having been tracked consistently, has not introduced any serious selection bias to our dataset.

5.2. Structure of trip purposes

The results of the classification of trip purposes and the assignment of routes to these purposes are presented and compared to the distribution of the MiD (2017) in Figure 3. Although the definitions of routes and categories are not identical in both figures, we observe similar patterns. In both samples, routes ending at the place of residence form the largest category, accounting for roughly one third of all trips. Our category "Commute" represents a similar share of routes as the respective MiD classes "Commute" and "Way to training location". However, the proportion of professional routes in our sample is much smaller than in the MiD data due to a lack of data forcing us to use a very narrow definition of such trips. The MiD definition of "Leisure" includes a subcategory "Visit of friends and relatives" corresponding to our category "Visit". This category is bigger in our sample with a share of 26% than in the MiD data, where it represents 18% of trips. The MiD category "Accompanying" can only be identified in our data in the specific situation of bringing children to school or kindergarten, which explains why the respective group is much smaller. An almost equal share of routes in both samples is conducted

Figure 3: Proportion of routes per trip purpose and comparison with the MiD (2017)



with the purpose of shopping. An issue arises in our identification of "Private Business" trips (e.g., visits to doctor, public offices, post office, etc.): as a fair proportion of those routes ends in points of industrial or commercial land use, they might be classified as "Undefined", "Professional", or "Shopping" (if they end close to a shopping amenity). The MiD classification does not contain a category corresponding to mode change or transfer, insofar as they should be part of a longer trip with a given purpose. We nonetheless use a transfer category, which exists in other forms of classification across the literature, for routes which end in transport hubs and seem to be part of longer multi-step trips. Finally, the share of undefined purposes is relatively small in both samples (under 10%), but a fundamental trade-off between the confidence level of the classification and the size of this category must be noted.[22]

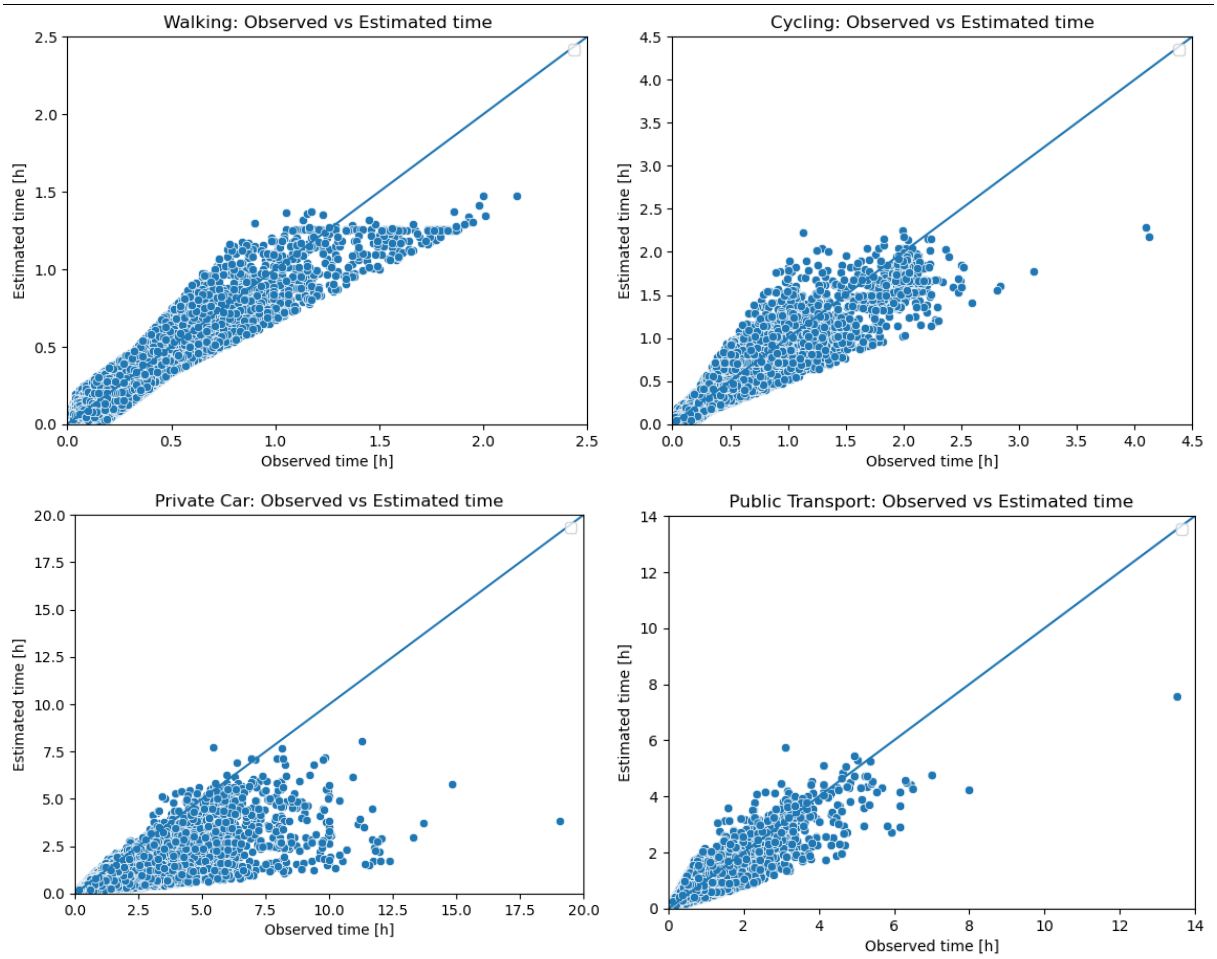
5.3. Bias between calculated and recorded trip duration

An important check for biased data is the comparison between the observed and the calculated trip duration. Figure 4 show a major underestimation of travel time for car routes (by almost one third when comparing mean durations) and a slight overestimation (around 4%) for public transport. In both cases, the differences between estimated and observed trip duration are larger for shorter trips. While the bias for car trips is expected due to ORS not considering congestion in estimated travel times, the slight overestimation of the trip duration calculated for public transport trips is surprising. Delays and cancellations have been a persistent problem in public transport and especially for trains throughout the observation period, which is only marginally reflected in the comparison between the observed trips and the data obtained from Google Maps and the DB website. However, since the 9-Euro ticket was not valid on long-distance trains, which experienced very low punctuality figures during summer 2022, the bias between calculated (scheduled) and observed travel times might be not as high as expected.

The extent to which travellers base their mode choice decision on such under- or overestimations remains unclear. Presumably, the decision-bias introduced by information drawn from sources like Google Maps is a problem for unknown routes, whereas decisions for repeated routes are primarily made using experience. The comparison shows that it is generally adequate to use the calculated travel times when modelling mode choice, as they correspond closely to the duration individuals consider when comparing different transport modes. However, it would be interesting to account for familiarity with routes (for example commute routes) to understand whether individuals base their decision on experience and assume higher travel

[22] It is possible to decrease the size of the "undefined" category by relaxing the criteria used for other categories. However, this reduces the level of confidence, i.e., the probability that the purpose of a route is assigned correctly.

Figure 4: Observed versus calculated trip duration per mode of transport



times than those given by Google Maps. If this is in fact the case, it might be appropriate to use models that include calculated travel times with a correction factor derived from this bias.

6. Conclusions

GPS-based data offers a wealth of relevant trip information, including precise start and end locations as well as corresponding start and end times, providing valuable insights for transportation analysis. When combined with surveys or socio-economic data of participants, collected at regular intervals, this type of information is capable of providing a more comprehensive image of travel behaviour than traditional methods such as travel diaries obtained from PAPI, CATI, or web-questionnaires. As mentioned above, trip distance and duration are recorded and therefore more reliable than self-reported values from travel diaries. In addition, trips of shorter distance (mainly walking and cycling), which are often under-reported in national travel surveys, are included.

However, tracking data also come with a cost. Despite the large amount of collected data, it is not necessarily complete, since people might (temporarily) uninstall/deactivate the tracking app or not take their smartphone with them, and devices might suffer from malfunctions or exhausted batteries. Additionally, geolocation tracking data can suffer from measurement errors, inconsistencies, and recording of routes that are not of interest for the envisaged analysis (for example circular routes or hiking routes). They therefore require extensive cleaning and processing efforts, coming along with a significant loss of raw data. Furthermore, the use of GPS data requires algorithms to assign the mode of transport, whereas traditional travel diaries include reported modes of transport. These algorithms have limited reliability especially when certain modes have similar speed profiles and use the same infrastructure, for example in the distinction between car, van, and truck, between bicycle and e-scooter, between own car, taxi, and Uber, or between types of trains. Other assignment problems refer to the treatment of congestion and, even more important, stop-and-go traffic: The stops might interrupt a recorded trip, requiring a re-combination of individual movements into routes, and the slow movements might be mis-identified as walking. Another disadvantage compared to travel diaries is the need to infer the trip purpose based on probabilistic models, machine learning methods, or rule-based approaches. If the data are collected to feed into a discrete choice model, trip costs have to be calculated and modal alternatives with their attributes have to be determined, as it is also the case when using reported trip data from travel diaries. Finally, the problem of multi-modal route choices is similar to travel diaries, as multimodal routes can be derived from recorded mobile phone data but can hardly be included in the alternatives required for a discrete choice model.

We conclude from our application that a dual approach of using both GPS-based data and survey information is preferable to either one of the two data sources alone as well as to more common measurements of travel behaviour. However, it is necessary that the survey contains not only the necessary socio-economic characteristics and basic information on travel behaviour, but also additional questions depending on the research question - in our application on the possession of certain types of tickets. This requires continuous participation of respondents in both the survey waves and the tracking, as the requirement of balanced panels reduces the final sample size considerably for many applications. The most important conclusion we draw is the need for a thorough check whether the necessary steps of data cleaning, combination, and processing introduce a bias in the remaining information, for example by affecting certain types of routes more than others.

References

Bansal, P., Hörcher, D., & Graham, D. J. (2022). A dynamic choice model to estimate the user cost of crowding with large scale transit data. *J. Roy. Stat. Soc.: Ser. A*, 2022, 1-25.

Bricka, S. G., et al. (2012). An analysis of the factors influencing differences in survey-reported and GPS-recorded trips. *Transportation research part C: emerging technologies,* 21.1: 67-88.

Calastri, C., et al. (2019). Mode choice with latent availability and consideration: Theory and a case study. *Transportation Research Part B: Methodological,* 123: 374-385.

De Grange, L., González, F., Muñoz, J. C., & Troncoso, R. (2013). Aggregate estimation of the price elasticity of demand for public transport in integrated fare systems: The case of Transantiago. *Transport Policy*, 29, 178-185.

Destatis. (2022). *9€ ticket no longer on offer - rail traffic back to pre-crisis level*. https://www.destatis.de/EN/Press/2022/09/PE22_377_12.html

Gaudry, M. J., & Dagenais, M. G. (1979). The dogit model. *Transportation Research Part B: Methodological,* 13.2: 105-111.

Gong, L., et al. (2014). Deriving personal trip data from GPS data: A literature review on the existing methodologies. *Procedia-Social and Behavioral Sciences,* 138: 557-565.

Hernández, D., & Witter, R. (2015). Perceived vs. actual distance to transit in Santiago, Chile. *Journal of Public Transportation,* 18.4: 16-30.

Link, H. (2015) Is car drivers' response to congestion charging schemes based on the correct perception of price signals? *Transportation Research Part A: Policy and Practice,* 71: 96-109.

Markttransparenzstelle für Kraftstoffe (MTS-K). (2021). *Jahresbericht 2021*. Bundeskartellamt. https://www.bundeskartellamt.de/SharedDocs/Publikation/DE/Berichte/Jahresbericht_MTS-K_2021.pdf

Martínez, F., Aguila, F., & Hurtubia, R. (2009). The constrained multinomial logit: A semi-compensatory choice model. *Transportation Research Part B: Methodological,* 43.3: 365-377.

MiD. (2017). *Mobilität in Deutschland 2017*. https://www.mobilitaet-in-deutschland.de/archive/publikationen2017.html. Accessed on 10 July 2023.

MOP. (2022). *Deutsches Mobilitätspanel. Eine Längsschnittstudie zum Mobilitätsverhalten der Bevölkerung*. https://mobilitaetspanel.ifv.kit.edu/index.php. Accessed on 10 July 2023.

Nguyen, M. H., et al. (2020). Reviewing trip purpose imputation in GPS-based travel surveys. *Journal of Traffic and Transportation Engineering (English Edition),* 7.4: 395-412.

Paul, K., et al. (2013). Quantifying the difference between self-reported and global positioning systems-measured journey durations: a systematic review. *Transport Reviews,* 33.4: 443-459.

Peer, S., et al. (2014). Over-reporting vs. overreacting: Commuters' perceptions of travel times. *Transportation Research Part A: Policy and Practice,* 69: 476-494.

Shen, L., & Stopher, P. R. (2014). Review of GPS travel survey and GPS data-processing methods. *Transport reviews,* 34.3: 316-334.

SOEP. (2022). *2022 Data for years 1984-2020, SOEP-Core v37, EU Edition*. doi:10.5684/soep.core.v37eu

Spurr, T., et al. (2015). A smart card transaction "travel diary" to assess the accuracy of the Montréal household travel survey. *Transportation Research Procedia,* 11: 350-364.

Tsoleridis, P., Choudhury, C. F., & Hess, S. (2022). Deriving transport appraisal values from emerging revealed preference data. *Transportation Research Part A: Policy and Practice,* 165: 225-245.

Varela, J. M. L., Börjesson, M., & Daly, A. (2018). Public transport: One mode or several? *Transportation Research Part A: Policy and Practice,* 113: 137-156.