Article

Using Probability Distributions for Projecting Changes in Travel Behavior

Andreas Radke *[ID] and Matthias Heinrichs [ID]

German Aerospace Center, Institute of Transport Research, 12489 Berlin, Germany; matthias.heinrichs@dlr.de
* Correspondence: andreas.radke@dlr.de; Tel.: +49-30-67055-9111

Abstract: Mobility is a must for human life on this planet, because important activities like working or shopping cannot be done from home for everyone. Present modes of transports contributes significantly to green house gas emissions while the efforts to reduce these emissions can be improved in many countries. Pathways to a more sustainable form of mobility can be modelled using travel demand models to aid decision makers. However, to project human behavior into the future one should analyze the changes in the past to understand the drivers in mobility change. Mobility surveys provide sets of activity diaries, which show changes in travel behavior over time. Those activity diaries are one of the inputs in activity-based demand generation models like travel activity pattern simulation (TAPAS). This paper shows a method of using probability distributions between person and diary groups. It offers an opportunity for an increased heterogeneity in travel behavior without sacrificing too much accuracy. Additionally it will present the use case of temporal back- and forecasting of changes in activity choices of existing mobility survey data. The results show the possibilities within this approach together with its limits and pitfalls.

Keywords: activity-based travel behavior; diary generation; TAPAS; travel demand generation; probability distribution

1. Introduction

Mobility is needed in any form of human society. Although the sectors of energy generation and food production are transforming towards sustainability for decades, the mobility sector in the industrialized countries is still heavily dependent on individual modes of transports, namely cars and trucks. In Germany, the total amount of carbon dioxide emissions in the transportation sector has not changed from 1990 to 2019, while most other sectors have significantly reduced their emissions [1]. To be able to form a sustainable future for the mobility sector decision makers have to know, how the usage of modes of transportation has changed and is changing. One way to support the decision makers to forecast travel demand by travel demand generation models, which simulate the demand of mobility for a given set of parameters in a specific region. The outcome of such models represents the reaction of the population to hypothetical introduced measures resulting in changes of trips and their length, performed activity, and used mode of transport. In case of car trips, these values can be used to calculate the CO_2 emission and traffic jams. A case of rebound effects regarding fuel efficiency improvements and the following increase in car traffic demand is shown and discussed in [2]. Because human behavior is not strictly logical or economical in every way these models have to deal with many uncertainties and cope with the errors statistically. Therefore, understanding why people leave their homes, how this behavior has changed over the years and how to forecast it is crucial for the quality of travel demand modeling to show pathways to a sustainable transport system with respect to the mobility needs. The *Mobilität in Deutschland* (MiD, Mobility in Germany) survey tracks mobility patterns of the German population since 2002 [3]. With its second [4] and third instalment [5] one can see a change in travel behavior

over time. A summary of the development of these surveys can be seen in the MiD time series report [6].

Figure 1 shows that from 2002 to 2017 the number of trips decreased with a slight bump in 2008. Looking at the specific activity categories the biggest changes are found in shopping, free time, personal matters and work. According to the MiD time series report [6] the increase in work-related trips is not because of more trips to work but rather business trips during work. One factor in the surge of business trips is assumed to be because of courier and parcel services. The MiD is unclear about the correlation between the reduction in shopping trips due to e-commerce (together with the increase in business trips) but states structural changes of shopping opportunities. Smaller stores like bakeries and butcher's shops disappeared or were integrated into bigger supermarkets which led to fewer trips ([6], pp. 59–62).
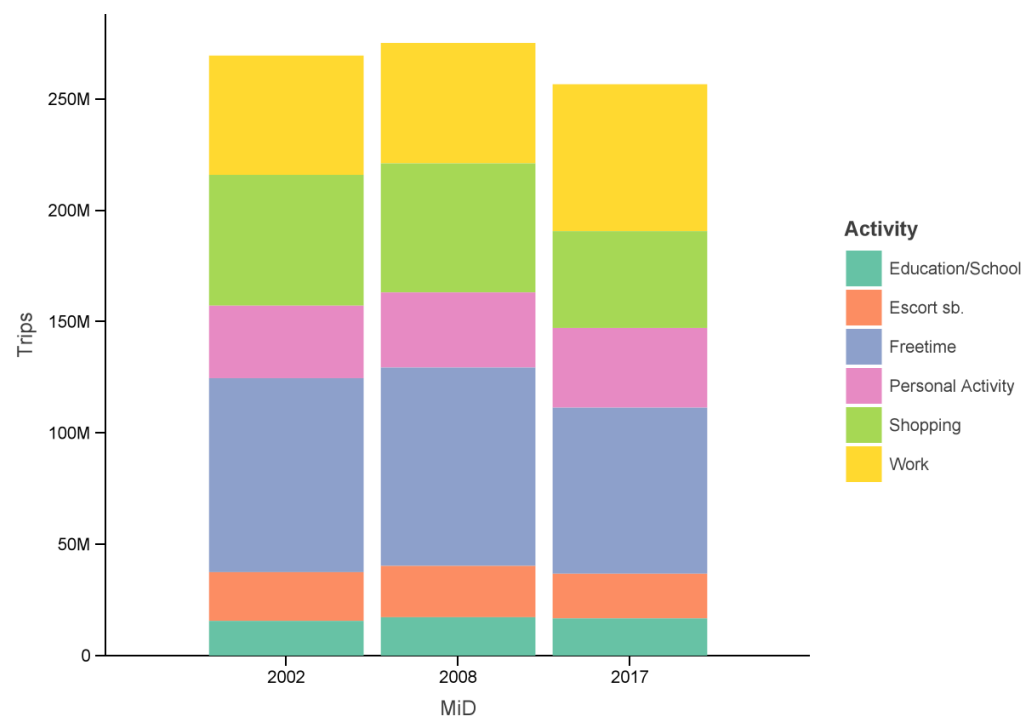
Figure 1. Number of trips per day ([6], compare p. 60 Figure 35).

Activity diaries play an important role in modeling travel behavior. Activity-based demand generation models like travel activity pattern simulation (TAPAS) [7–10] rely on large datasets of mobility surveys such as the MiD. In general, travel demand models separate their simulation into four sub-models: First, trip generation, where private trips with specific purposes are generated due to their socio-economic status, age, employment, etc. Second, trip distribution: here a location choice for the desired trip purpose is done based on accessibility, type of location, personal constraints, capacity, and occupancy. The third step is the choice of mode of transportation, which is often integrated in location choice, because one cannot exist without the other. The mode choice highly depends on availability of the mode, pricing, travel time, and trip purpose. The fourth and last step is called traffic assignment, where the routes for each trip is calculated and capacity constraints of the roads and buses are taken into account.

In this paper, we present an enhanced method using probability distributions based on trip purpose hierarchies. This addresses the first step in the previously described four-step-model. A similar approach was shown in [10–13] but with a different set of group divisions. Furthermore, this previous work could not analyze the performance of the proposed approach due to the lack of consecutive installments of their primary data source. We test the accuracy of this concept while obtaining more heterogeneity in modeled

travel behavior. Because it is extremely expensive to do these mobility surveys on a scale which produces meaningful data for demand generation models, we will show a way of keeping existing diaries of different years. In a method of fore- and backcasting, we will display an opportunity to allow and reflect individual travel behavior changes over time only by changing the probabilities without a change in the used diaries or doing a complete resurvey of mobility patterns. Resulting, it will be shown how well they can be projected into the future by forecasting later MiDs from the past data and the previous ones from the recent one.

This paper is organised as follows: At first we will give a rough overview of the used underlying mobility survey and population data. After an extension of these data with a new weighting of person groups, we will introduce our diary classification and its connection to person groups and probability distributions. The results Section 3 presents the computation of the distribution of activities with our probability distribution approach and compares it to the values from the MiD. Section 4 discusses the results, displays the approach's limits and offer further points of future research.

2. Materials and Methods

2.1. Data

The reported diaries are taken from the *Mobilität in Deutschland* (MiD, Mobility in Germany) time series report [6]. This report is an adaptation of the MiD surveys of 2002 [3], 2008 [4] and 2017 [5] with data weight recalibration for better comparison. Some diaries were removed during preprocessing because of missing necessary information. The first two surveys were vastly smaller than the one from 2017 with the latter roughly five times bigger (see Table 1).

Table 1. Number of diaries and trips reported in each MiD survey.

MiD	Diaries	Trips
2002	43,876	160,011
2008	49,591	179,486
2017	239,503	835,805

Each detailed diary report d consists of a set of trips $t_1, t_2, \ldots t_{n_d}$ with $n_d \in \mathbb{N}$. Table 2 presents exemplary data for two typical diaries in the MiD.

Table 2. Exemplary MiD diary data with attributes of most significance for this paper.

Person	1	1	1	2	...
Trip	1	2	3	1	...
Activity	free time	work	free time	shopping	...
Age	26	26	26	41	...
Status	student	student	student	not working	...
Sex	f	f	f	m	...
Cars	0	0	0	1	...
Weekday	Wed	Wed	Wed	Sat	...
Start Time	7:30	9:00	18:00	10:00	...
End Time	7:45	9:30	18:30	10:10	...
Region	≥0.5 M inh.	≥0.5 M inh.	≥0.5 M inh.	<5000 inh.	...
Diary Group	5	5	5	20	...
Person Group	Student	Student	Student	Not working, ≥25, <45, m, w/Car	...

Out of the many attributes of each trip the following attributes are of major interest to our research: Person attributes, such as

- age;
- working, educational or retiree status;
- sex; and
- car-ownership.

For the filtering the dataset to investigate more specific cases, e.g., a typical work day in bigger cities:

- The weekday;
- Region class (metropolitan, rural area etc.), i.e., where the person lives.

Additionally, for categorizing the diaries we use

- The activity; and
- The start and end time of the trips to distinguish between full time and part time work.

This gives us for each diary d:

- A set of trips, we write for the trips $t_1, t_2, \ldots, t_{n_d} \in d, n_d \in \mathbb{N}$;
- An activity of a trip t is denoted by a or $a(t)$ with

  $a \in A := \{\text{Any Activity, Education/School, Free Time, Personal Activity, Shopping, Work}\}$;

- $g_d(d)$ denotes the diary group of diary d; and
- $g_p(d)$ denotes the person group to which the reporting person of diary d belongs.

The set of diaries is denoted by D. G_d is the set of diary groups and G_p is the set of person groups. Where applicable, we write D_{2002}, D_{2008}, D_{2017} for the diary sets of the respective MiD survey or D and D_x for diaries without specifying the year.

We used a synthetic population of Berlin with 3,604,320 citizens created by Synthesizer [14]— an internal tool from the DLR Transport Research institute. The population is based on data from the *Mikrozensus* 2015 [15], *Zahlen–Daten–Fakten Berufliche Schulen from the Senatsverwaltung für Bildung, Jugend, Familie Berlin* [16], Nexiga Demography and Households 2017 [17,18], *Rentenatlas* 2018 [19], and *Statistisches Jahrbuch Berlin-Brandenburg* [20].

*2.2. Synthetic Population and Weighting*

There are two main reasons for changes in travel behavior (compare [11]):

1. Changes in the population, i.e., increase in younger or older people, changes in employment etc.; and
2. Changes in individual travel behavior like working less, having more free time or e-commerce replacing some amount of shopping trips.

We want to remove the first reason from our consideration and only investigate the changes in individual travel behavior. For a better representation of the German population, the MiD assigns a weight to each diary according to the attributes of the reporting person. These weights correspond to the respective population of the years of the MiD survey (2002, 2008, and 2017). To remove the differences in trips and activities due to population changes, we give each MiD a new set of weights and ignore the weights from the MiD altogether. For the new weighting we use our synthetic population of Berlin with roughly 3.6 million people. Each person belongs to one of 34 person groups.

Figure 2 details the distribution of the person groups. The status like student, working or pupil defines the first split into several group segments. The numbers specify the age range, e.g., from 25 to under 45. The sex is stated by male or female. For groups where the gender is not of importance we write m/f or omit it entirely. "*W/(o) car*" indicates the car ownership. Sex, age, and car attributes are omitted if it is not of significance for the person group, like pupils, students, and trainees. We decided on this group division by analyzing the available input data and forming homogeneous user groups of interest. The age classes

are chosen to reflect certain periods of life, like first job (<25), young professionals (25–45), senior professionals (45–65), young retirees (65–75) and old retirees (>75). Doing so we had to meet two external constraints: The group size must not drop to less than 100 diaries. Separation between male and female is only necessary, if the frequency of activities in their diaries differ more than 1%.
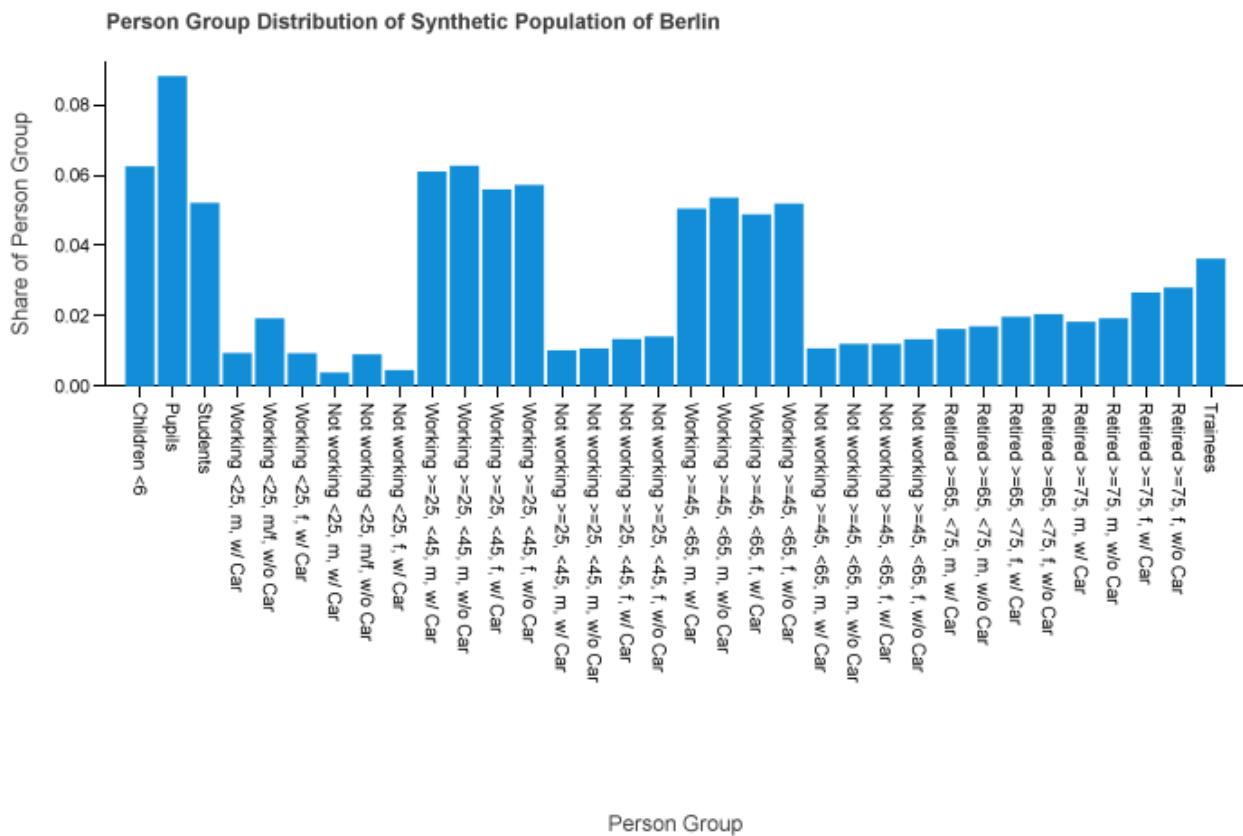
Figure 2. Person group distribution of a synthetic population of Berlin.

One can immediately see the smaller share of unemployed people or the higher share of retired women compared to men. The attributes from Section 2.1 are used to classify a person to a person group like students or working women between 45 and (excluding) 65 without a car. Therefore (as seen as in Table 2), each diary d is assigned to a person group. We write $g_p(d)$ or only g_p if it is obvious to which diary it refers to or if the diary is not of importance. For the weighting we considered two diary filters:

1. Diaries of all regions from Monday to Sunday; and
2. Diaries from regions with more than 0.5 M inhabitants during core weekdays (Tuesday to Thursday) only.

Filter 1 represents an average of the whole week over all regions but with the population distribution of Berlin. This may be no realistic image but suffices for research purposes. For Filter 1, we write the set of diaries as D_x. Filter 2 models a typical workday in a metropolitan area. This case may be transferable to other larger cities in Germany, such as Hamburg, Munich or Cologne, but further attention to the respective population is needed. For Filter 2, we write $\bar{D}_x$.

To adapt the diaries we gave each diary d with person group $g_p(d)$ a weight $w(d)$ depending on the filter. Because each diary within one person group and MiD set will have

the same weight, we can write $w(d) = w(g_p(d)) = w(g_p)$. The weight of a specific person group g'_p is defined as

$$w(g'_p) := \frac{|D|}{|\{d \in D | g_p(d) = g'_p\}|} \cdot \frac{|g'_p|}{popul(Berlin)}$$

where $|g_p|$ is the shorthand notation of the number of people in person group g_p from the synthetic population. D is a placeholder for D_x and $\bar{D}_x$, where $x \in \{2002, 2008, 2017\}$. $popul(Berlin)$ is the number of people in Berlin. $|X|$ denotes the cardinality of set X as usual.

### 2.3. Diary Classes and Probability Distributions

For a microscopic agent-based simulation of the activity travel behavior of a population one could be satisfied with a single division into person groups. Each person (i.e., agent) takes a reported diary of its person group. This leads to problems due to a lack of reported diaries in specific person groups. For example the person group of non-working under 25-year-olds without a car reported only 39, 32, and 56 diaries in 2002, 2008, and 2017, respectively. This further decreases if someone wants to use the diaries of metropolitan regions during the middle week days to get an image of a typical work day in bigger cities like Berlin. In this case only 3, 5, and 6 diaries are reported respectively.

Because of this, we use diary groups which assign to each diary a specific group with a special commonality between all diaries within a group. Hertkorn et al. [12,13] uses sequence alignment and clustering algorithms to classify diary groups. We use a different and simpler classification of diary groups with a more straight forward way of assigning the diaries by its activities. We discuss the uncertainties of this approach later in Section 4. Table 3 presents the diary groups we used. Note that, despite escort trips defining some diary groups we will later conflate escort trips into any activity.
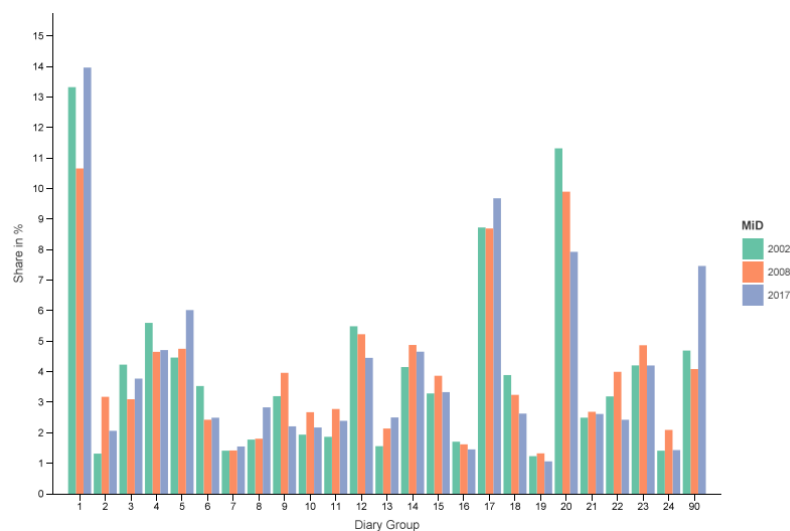
Table 3. Diary groups.

Diary Group Number	Diary Group
(1)	Full time work trip
(2)	Full time work with escort trip,
(3)	Full time work with personal matter trip,
(4)	Full time work with shopping trip
(5)	Full time work with free time trip
(6)	Part time work trip
(7)	Part time work with escort trip
(8)	Part time work with personal matter trip
(9)	Part time work with shopping trip
(10)	Part time work with free time trip
(11)	Educational trip for students
(12)	Educational trip for pupils
(13)	Kindergarten/Educational trip for children <6
(14)	With Escort trip
(15)	Personal matter trip for full/part-time worker
(16)	Personal matter trip, students, pupils, children
(17)	Personal matter trip, non-working, retirees
(18)	Shopping trip for full/part-time worker
(19)	Shopping trip, students, pupils, children
(20)	Shopping trip, non-working, retirees
(21)	Free time trip for full/part-time worker
(22)	Free time trip, students, pupils, children
(23)	Free time trip, non-working, retirees
(24)	Educational trip for Trainees
(90)	Other diaries

The distribution for all regions and all days can be seen in Figure 3a and for metropolitan regions from Tuesday to Thursday in Figure 3b. One can see that share of working (1–10) and educational (11–13, 24) diaries is higher for the core weekdays. Furthermore, a drastic decrease in free time diaries of persons who (usually) work (21) or go to an educational institution (22) with a smaller reduction for non-working people (23) is visible when comparing the two filters.
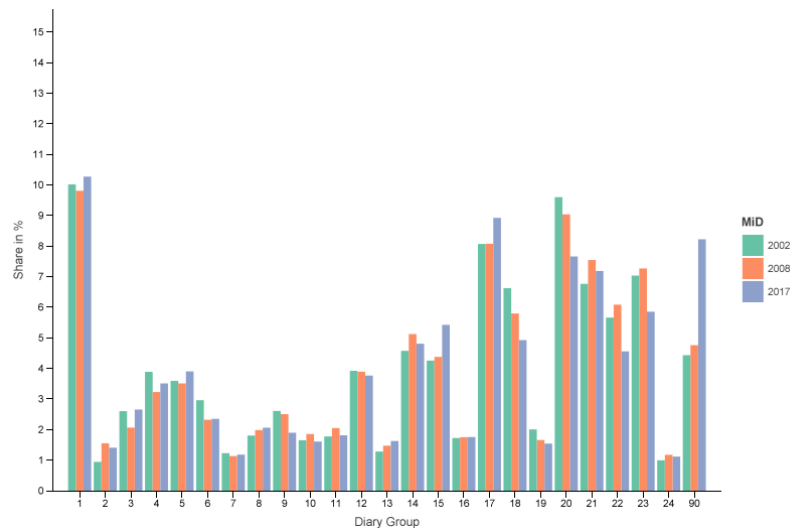
(**a**)

(**b**)

Figure 3. Diary group distribution. (**a**) D_x, all Regions from Monday to Sunday. (**b**) $\bar{D}_x$, cities with $\geq$0.5 M inhabitants from Tuesday to Thursday.

Note that, the diary groups have in our case different priorities. The highest priorities are educational diaries for specific groups like children, pupils, students and trainees. Educational trips by working or non-working people—such as going to a language class in the evening—are of lowest priority. Other than that it goes roughly in the order of its numbering. The exceptions are

- (1) which comes after (2), (3), (4), (5); and
- (6) which comes after (7), (8), (9), (10)

For example if a diary of a student is reporting a trip to the university it will belong to diary group (11) no matter if the student is going to do its students job on the same day.

Another example: If an employee is not going to work on the day of the report but goes shopping and to their yoga class (free time), the diary will belong to group (18), but not (21), because (18) has the higher priority over (21).

The purpose of these diary groups is to have a greater pool of available diaries for the person groups and introduce less homogeneity. Considering the example of a student, it may be the case that the student is going to the university and, hence, diary group (11) is chosen. Nevertheless, a student may behave on a single day like a typical full time worker and doing their 8-hour shift of their students job. In case of a full-time work day the person chooses a diary of group (1)–(5) or for part-time work group (6)–(10).

This leads to a probability distribution where $prob(g_d|g_p)$ denotes the probability of a person in person group g_p choosing a diary in diary group g_d.

$$prob(g'_d|g'_p) = \frac{|\{d \in D|g_p(d) = g'_p \wedge g_d(d) = g'_d\}|}{|\{d \in D|g_p(d) = g'_p\}|} \tag{1}$$

where D is again a placeholder for D_x and $\bar{D}_x$. Furthermore, we specifically write $prob_x(g_d|g_p)$ and $\overline{prob}_x(g_d|g_p)$. It holds

$$\sum_{g_d \in G_d} prob(g_d|g_p) = 1$$

where G_d ($= \{(1), \ldots, (24), (90)\}$ in our case) is the set of diary groups. For a person p of group g_p, instead of only using the diaries belonging to the person group g_p this approach enables us to possibly assign any diary in the diary groups with $prob(g_d|g_p) > 0$.

3. Results

3.1. MiD Data Results

To get the share for each activity $a \in A$ we compute

$$share_{MiD}(a) = \frac{\sum_{d \in D} \sum_{\substack{t \in d \\ a(t)=a}} w(g_p(d(t)))}{\sum_{d \in D} \sum_{t \in d} w(g_p(d(t)))},$$

where $w(g_p(d(t)))$ is the weight of person group $g_p(d(t))$ of diary $d(t)$ of trip t.

The specific values in percent for the two filter sets Filter 1 (D_x) and 2 ($\bar{D}_x$) of $share_{MiD}(a)$ can be seen in Table 4 and in Figure 4a,b.

Table 4. Activity share of the MiD time series report for the two filters.

	All Regions, Mon–Sun			≥0.5 M, Tue–Thu		
	2002	**2008**	**2017**	**2002**	**2008**	**2017**
Any Activity	7.58	9.06	7.85	7.91	9.97	8.50
Education/School	5.42	6.05	6.43	6.92	8.08	8.20
Free Time	34.68	34.83	35.25	27.16	27.10	28.59
Personal Activity	12.87	13.00	15.07	13.63	13.45	14.97
Shopping	22.50	20.60	18.00	23.27	20.90	16.93
Work	16.96	16.47	17.40	21.11	20.50	22.81
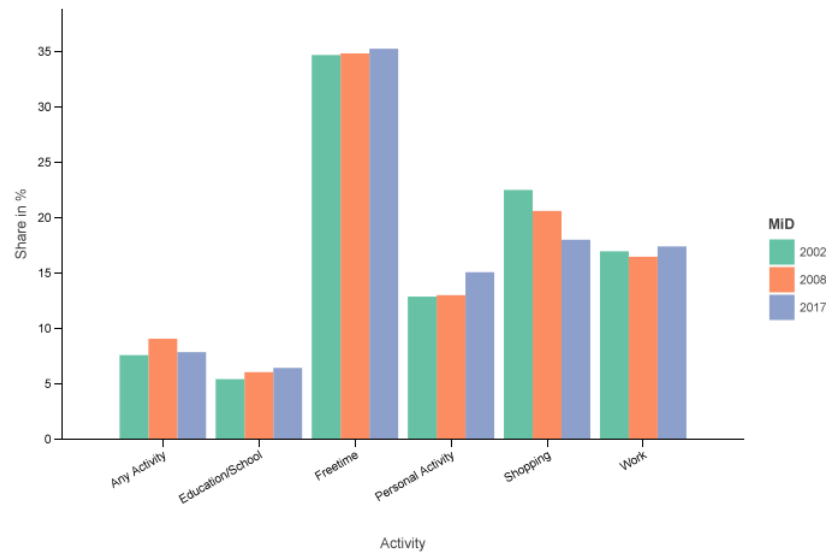
(**a**)
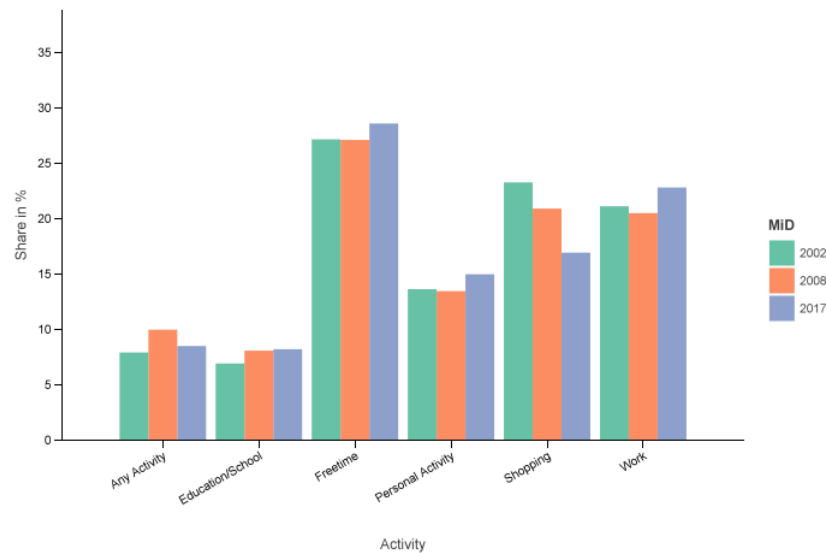
(**b**)

Figure 4. Activity share in the MiD over time. (**a**) All regions, Monday–Sunday. (**b**) Cities with more than 0.5 M people, Tuesday–Thursday.

Free time has the biggest share in all cases by a good margin. Unsurprisingly, one can see a lower share of roughly eight percentage points for free time in the core weekdays in metropolitan areas. The shopping activities, especially in Figure 4b, have the biggest decrease over time. All other activities are at least in relative parts increasing from 2002 to 2017, either strictly (e.g., education) or with a bump or dent in its course (e.g., any activity, work).

3.2. Diary Class Distribution Results

When using the probability distribution from Equation (1), we compute the $share_{pd}(a)$ of each activity a for our Berlin population applied to each MiD report through

$$share_{pd}(a) = \frac{\sum_{d \in D} \sum_{\substack{t \in d \\ a(t)=a}} \sum_{g_p \in G_p} prob(g_d(d(t))|g_p) \frac{|g_p|}{|g_d|}}{\sum_{d \in D} \sum_{t \in d} \sum_{g_p \in G_p} prob(g_d(d(t))|g_p) \frac{|g_p|}{|g_d|}}. \tag{2}$$

Like above, $|g_p|$ is the cardinality of the person group in the Berlin population. $|g_d|$ denotes the number of diaries in diary group g_d with respect to D_x. Note that, to maintain diary heterogeneity, we use the whole set of diaries of each year $\{D_{2002}, D_{2008}, D_{2017}\}$, respectively. In case of the diary class distributions the Filter 2 is only applied to the probability computation.

Table 5 shows the distribution of activities for each combination

$$x, y \in \{2002, 2008, 2017\} \times \{2002, 2008, 2017\}$$

of diaries $D_x, \bar{D}_x$ and probability distributions $prob_y, \overline{prob}_y$. Comparing the MiD 2002, 2008, 2017 development of the frequencies of the activities to the entries where only the probability distributions $prob_y, \overline{prob}_y$ change one can make the following observations:

- The development of free time and any activity is never met;
  - These two activities have the lowest priority according to our diary group order;
- The development of education is only achieved for all regions and all days (it rises twice from 2002 to 2017) but not for bigger cities in the core workdays (again rises twice). The probability prediction states an increase at first and a smaller decrease from 2008 to 2017;
  - A (half) mis-prediction despite being the highest priority group for children, pupils, students, and trainees;
  - The share of educational trips in diaries not in diary class (11), (12), (13), (24) are 1.28% (2002), 1.15% (2008) and 1.17% (2017) for all regions and days. Because of these diaries which can fall into any diary group the number of estimated educational trips seems always lower than the reported number from the MiD;
  - For bigger cities, Tuesday–Thursday we have 1.68%, 1.56% and 2.26%. This also leads to an increased error (again using all diaries but with probabilities from the filtered set), especially for the 2017 set;
- The trends of work, personal matters and shopping are reached in both cases.

Table 5. Overview of activity percentages for each combination of diary and probability distribution. Columns where the years of diaries correspond to the probability distribution are highlighted. $D_x, prob_y$ and $\bar{D}_x, \overline{prob}_y$ are the respective filters 1 and 2.

Diaries D_x	2002			2008			2017		
Distribution $prob_y$	2002	2008	2017	2002	2008	2017	2002	2008	2017
ine Any Activity	7.51	8.37	9.46	7.96	8.86	9.96	6.46	7.28	7.58
Education/School	5.25	5.60	5.85	5.10	5.46	5.71	5.22	5.58	5.77
Free Time	35.19	35.94	33.54	34.77	35.47	33.04	36.17	37.00	35.35
Personal Activity	12.93	12.97	14.50	12.83	12.88	14.41	13.60	13.68	15.31
Shopping	22.12	20.89	19.87	22.29	21.05	20.03	20.64	19.38	18.30
Work	17.00	16.23	16.78	17.05	16.30	16.84	17.91	17.08	17.70
ine Diaries $\bar{D}_x$		2002			2008			2017	
Distribution $\overline{prob}_y$	2002	2008	2017	2002	2008	2017	2002	2008	2017
ine Any Activity	7.44	8.72	9.43	7.87	9.23	9.94	6.43	7.96	7.86
Education/School	6.22	6.95	6.74	6.07	6.79	6.60	6.29	7.07	6.76
Free Time	28.67	30.00	28.08	28.48	29.68	27.76	29.47	30.56	29.46
Personal Activity	13.55	13.23	14.50	13.41	13.10	14.33	14.18	13.91	15.22
Shopping	22.28	21.02	19.03	22.30	21.02	19.08	20.64	19.36	17.31
Work	21.83	20.08	22.22	21.86	20.17	22.28	22.99	21.14	23.39

Figure 5 presents the difference between Tables 4 and 5. In more detail, the figure shows the result of $share_{pd}(a) - share_{MiD}(a)$ with $share_{pd}$ in respect to $D_x, prob_y$ and

$share_{MiD}$ in regard to diaries D_y. The same is done for Filter 2 and the corresponding sets. In the x axis, the first number corresponds to year x and the second to year y. The maximal absolute error in (2002, 2002), (2008, 2008), (2017, 2017) is below one percentage point for Filter 1 and increases to 2.58 percentage points difference for Filter 2. The bigger error for the latter one may be because of the used diaries of all regions and days which are not representative of bigger cities and core workdays. One example causing this effect might be having less free time in the core week. Especially for the columns where the year of the used diaries coincides with the year of the used probability distributions have very low differences. Considering all columns the absolute maximum error increases to 2.21 and 3.45 percentage points, respectively.
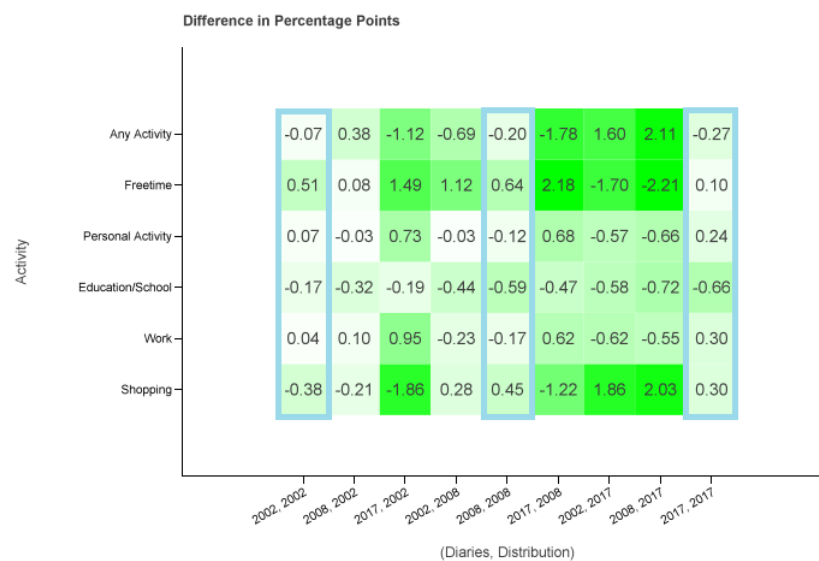
(**a**)
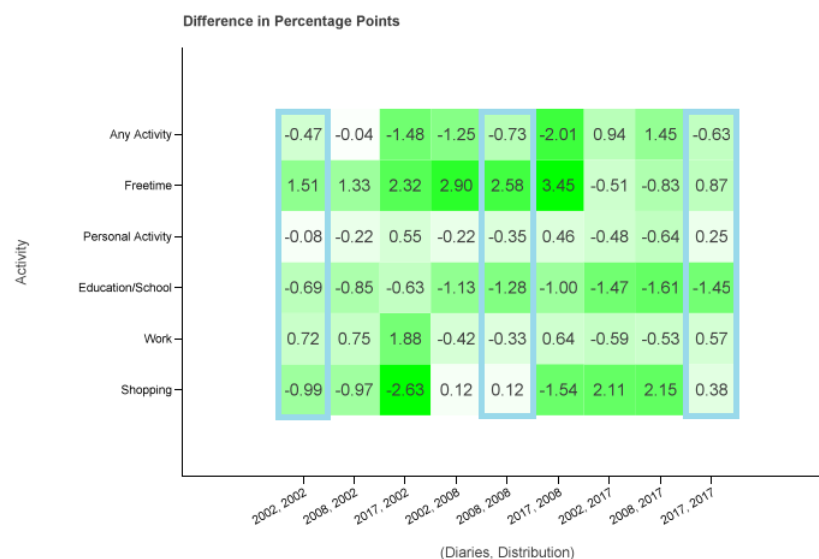
(**b**)

Figure 5. Difference in percentage points. The differences are taken against the MiD values of the same year as the probability distributions. (**a**) All regions, Monday to Sunday (**b**) Cities $\geq$0.5 M, Tuesday to Thursday.

The absolute value of differences is particularly small for the activities education, work and personal matters regardless the combination of x and y. The absolute difference for these activities is never above 1.0 ($D_x, prob_y$) or 2.0 ($\bar{D}_x, \overline{prob}_y$) as opposed to the other

three activity categories. The effect of activity priorities in building the diary groups seems to be substantial.

3.3. Union of Diaries

To further increase the available set of diaries we now use the whole MiD time series report set

$$D_{all} = \{D_{2002} \cup D_{2008} \cup D_{2017}\}. \tag{3}$$

The probability distributions remain unchanged and specific to a single survey year, as seen in Equation (1). The activity shares of the respective year and filter are displayed in Table 6. Looking at the development of activity behavior again one can see a similar picture compared to the data from Table 5 with personal activities, shopping, and work matching the trend, free time, and any activity failing and educational trips doing both with the corresponding filter.

Table 6. Activity shares using the diaries D_{all} of 2002, 2008 and 2017 together with the probability distributions of a single year.

Filter	All Regions, Mon–Sun			$\geq$0.5 M, Tue–Thu		
Distribution $prob_y$	2002	2008	2017	2002	2008	2017
Any Activity	6.76	7.59	8.03	6.70	8.19	8.24
Education/School	5.18	5.54	5.74	6.22	6.99	6.70
Freetime	35.93	36.74	34.95	29.33	30.45	29.20
Personal Activity	13.42	13.48	15.09	14.00	13.72	15.02
Shopping	21.06	19.81	18.74	21.08	19.81	17.78
Work	17.66	16.85	17.45	22.67	20.85	23.07

The closest results delivers $(D_{all}, prob_{2017})$ compared to the MiD 2017 values (see Figure 6). It is only surpassed by $(D_{2017}, prob_{2017})$ for the same MiD from Figure 5. Given that the 2017 diary set outnumbers the sets from 2002 and 2008 drastically, $(D_{all}, prob_{2002})$ and $(D_{all}, prob_{2008})$ yield a less accurate outcome. Once again the projection models educational, work, and personal matter trips more accurate similar to Section 3.2.


(a)

Figure 6. *Cont.*
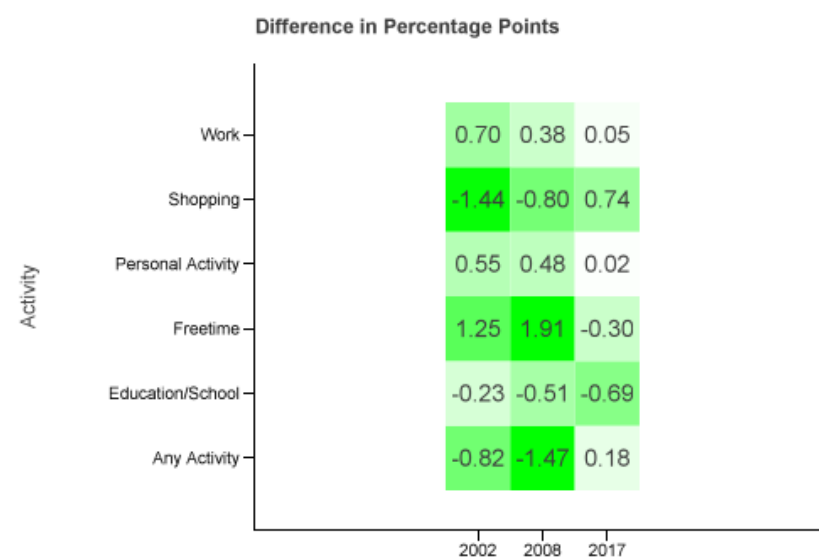
Difference in Percentage Points
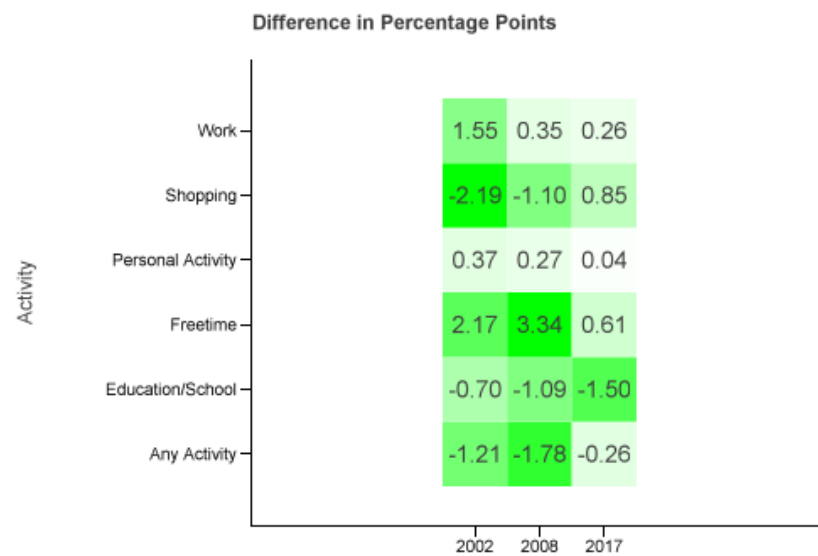
(**b**)

Figure 6. Difference in percentage points. Probability distributions of each survey combined with the whole set of diaries D_{all} of the three reports (2002, 2008, 2017). The differences are taken against the MiD values of the same year as the probability distributions. (**a**) All regions, Monday to Sunday. (**b**) Cities $\geq$0.5 M, Tuesday to Thursday.

We see that the $\overline{prob}_y$ differences are greater than for $prob_y$. To verify that the increased error can be reduced by using only diaries of bigger cities from Tuesday to Thursday we consider

$$\bar{D}_{all} = \{\bar{D}_{2002} \cup \bar{D}_{2008} \cup \bar{D}_{2017}\}.$$

Figure 7 displays the differences of the combinations of $\bar{D}_{all}$, $\overline{prob}_y$. Comparing it to Figure 6b one can see that the absolute maximum error decreases from 3.34 to 2.68 percentage points. In fact, there are two cells (shopping and personal matters in 2017) with a bigger and one cell (shopping in 2002) with the same absolute difference. All other activities and years are closer to the MiD data.

Difference in Percentage Points
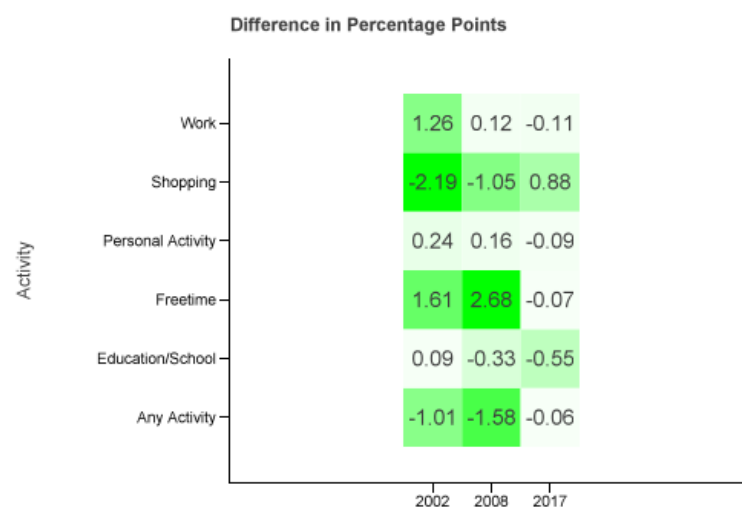
Figure 7. $\bar{D}_{all}$ only using diaries from cities $\geq$0.5 M from Tuesday to Thursday.

The filtering of used diaries increases accuracy. There are 24,056 diaries in $\bar{D}_{all}$ as opposed to 332,970 in D_{all}. The question remains if there are sufficient diverse diary data for the individual use case.

4. Discussion

This paper shows a way to use probability distributions between person and diary groups which enables the use of more diaries per person group while still being fairly accurate.

Even though we did not change the population in our computation, this approach is sensitive to changes in person group distributions. A calibrated (i.e., weighted) set of diaries would still work in a new and possibly future population. An increase in, for example, pupils and students would increase the number of educational trips. With new probability distributions, possibly derived from less extensive additional mobility behavior surveys, one could use this approach and project the future activity choices based on the survey results.

Trends like online shopping, part-time or remote work had an impact on the mobility behavior in the past and is expected to have impact in the future too. However, in the past, trips which became unnecessary were not omitted but are rather replaced especially by free time activities. A change in the number and purpose of trips strongly affects mode choice and, hence, air pollution or CO_2 and noise emission, especially considering trips by cars. The projection of these changes in travel behavior together with a demographic change need to be considered for developing evaluation strategies and political measures towards a sustainable mobility. A similar reasoning and its connection between the mobility behavior and the environment is explained in the DLR project report of *Transport and the Environment* (VEU) [21].

The findings of this work are included in the travel demand generation program TAPAS [7,8] from the Institute of Transport Research of the DLR. The first step in TAPAS is to assign activities for each person in the study area, using the probability distributions from this work. Afterwards the locations and modes of transport are chosen with respect to the personal mobility options and the spatial constraints like public transport service or access restriction for cars. Finally, the resulting diary plans are evaluated by their financial and temporal feasibility. This process is repeated until a feasible diary plan for every person is found. Doing so makes the final plan sensitive to changes in activities, locations, and modes. A full simulation output of a study area represents the decisions of the population with respect to the simulated political measures. Again, the probability distributions lead a way of modeling activity behavior. However, it is also possible that under a given activity behavior and new political measures (e.g., area restriction, gasoline price increase) the simulated outcome of chosen activities can differ from the national households surveys due to many retries in the diary plan selection. As a result we can measure if the desired sustainability goals are reached in the simulation and how the population has to adapt its activities via the presented probability distributions to the simulated scenario.

In parts future (or past), travel behavior can be projected with our model. Depending on the activity and its priority, like work and personal matters, the development is depicted accordingly in an accurate way. Other activities, such as free time, exhibit more inexact results, have outliers or show wrong trends.

The use of a more filtered dataset, such as the metropolitan regions from Tuesday to Thursday (Filter 1, $\bar{D}$, $\overline{prob}$) leads to less accuracy. The quandary is to only use the diaries from the filter (not only for the probability distribution) and be more accurate versus using all diaries and have a more diverse behavior.

Problems Needing Further Investigation

One important question is, is the classification of person groups and diaries reasonable and the most accurate way? It might be the case that car ownership is of importance for the mode choice but not for the activity choice. It seems plausible that a person's or household's activity choices are more affected by having one or more children leading to more escort trips for instance. A deeper analysis of persons and household attributes and the resulting choices may lead to different groups. Another option would be the use of clustering algorithms, such as Hertkorn et al. [12,13] and Varschen [11] demonstrated for person groups. Nevertheless, the remaining obstacles could be the justification, the lack

of comprehensibility of clustered groups, and the transferability to other mobility and household surveys.

We have seen that the priority in the assignment of diaries to the diary groups play an important role in the accuracy and quality of mimicking the trends in travel behavior. Presumably the current diary group division is not the best one possible. Especially, because none of the probability time series were capable of following the trend of free time trips despite free time contributing the largest amount of trips. A reconsidering of priorities is compelling. Hertkorn [13] and Varschen et al. [11] used sequence alignment and clustering algorithms to classify diaries into diary groups. It is not clear to what extent a general overhaul of the diary classification similar to the person groups may be necessary, but it is an interesting research topic nevertheless.

The two divisions of persons and diaries together may lead to small numbers of reports. This could cause over- or under-representation of specific behavior. Larger mobility datasets and more diaries may overcome this problem but mobility surveys are expensive. For now, it does not seem realistic to have an improvement in this data situation. One option may be the generation of synthetic diaries. Given that these are sufficiently realistic and accurate, a large number would further increase the individual heterogeneity even with or without an extra diary class distribution.

Further research in these areas are necessary in the future to improve the methodology of probability distributions in travel demand generation.

Abbreviations

The following abbreviations are used in this manuscript:

MiD	*Mobilität in Deutschland* (Mobility in Germany) survey
d	Diary
D_x	Diaries of all regions and all weekdays of year x
$\bar{D}_x$	Diaries of regions with more than 0.5 M inhabitants from Tuesday to Thursday of year x
g_d	Diary group
g_p	Person group
G_d	Set of all diary groups
G_p	Set of all person groups
$prob_y$	Probability distribution with respect to diaries D_y
$\overline{prob}_y$	Probability distribution with respect to diaries $\bar{D}_y$
$|X|$	Number of elements in set X.

References

1. Umweltbundesamt 2021, Treibhausgasemissionen: Emissionsquellen. Available online: https://www.umweltbundesamt.de/themen/klima-energie/treibhausgas-emissionen/emissionsquellen#energie-stationar (accessed on 9 August 2021).
2. Höltl, A.; Heinrichs, M.; Macharis, C. Analysis of rebound effects resulting from improved vehicle efficiency applied to the Berlin city network. In *Sustainable Urban Transport*; Attard, M., Shiftan, Y., Eds.; Emerald Group Publishing: Bingley, UK, 2014; Volume 7, pp. 229–249; ISBN 978-1-78441-616-4.
3. Institut für Angewandte Sozialwissenschaft (infas); Deutsches Institut für Wirtschaftsforschung (DIW). Mobilität in Deutschland 2002—Ergebnisbericht, Studie von infas, DIW im Auftrag des Bundesministeriums für Verkehr und Digitale Infrastruktur. Available online: http://www.mobilitaet-in-deutschland.de/pdf/ergebnisbericht_mid_ende_144_punkte.pdf (accessed on 11 August 2021).
4. Institut für Angewandte Sozialwissenschaft (infas); Deutsches Zentrum für Luft- und Raumfahrt (DLR). Mobilität in Deutschland—MiD 2008 Ergebnisbericht, Studie von infas und DLR im Auftrag des Bundesministeriums für Verkehr und Digitale Infrastruktur. Available online: http://www.mobilitaet-in-deutschland.de/pdf/MiD2008_Abschlussbericht_I.pdf (accessed on 11 August 2021).
5. Nobis, C.; Kuhnimhof, T. Mobilität in Deutschland—MiD 2017 Ergebnisbericht, Studie von infas, DLR, IVT und infas 360 im Auftrag des Bundesministeriums für Verkehr und Digitale Infrastruktur. Available online: http://www.mobilitaet-in-deutschland.de/pdf/MiD2017_Ergebnisbericht.pdf (accessed on 11 August 2021).
6. Nobis, C.; Kuhnimhof, T. Follmer, R.; Bäumer, M. Mobilität in Deutschland—Zeitreihenbericht 2002–2008–2017. Studie von infas, DLR, IVT und infas 360 im Auftrag des Bundesministeriums für Verkehr und Digitale Infrastruktur. 2019. Available online: http://www.mobilitaet-in-deutschland.de/pdf/MiD2017_Zeitreihenbericht_2002_2008_2017.pdf (accessed on 11 August 2021).
7. DLR Institute of Transport Research, TAPAS Webpage. Available online: https://www.dlr.de/vf/desktopdefault.aspx/tabid-12751/22270_read-29381/ (accessed on 11 August 2021).
8. DLR Institute of Transport Research, TAPAS Source Code. Available online: https://github.com/DLR-VF/TAPAS (accessed on 11 August 2021).
9. Heinrichs, M.; Krajzewicz, D.; Cyganski, R.; von Schmidt, A. Introduction of car sharing into existing car fleets in microscopic travel demand modelling. *Pers. Ubiquitous Comput.* **2017**, *21*, 1–11. [CrossRef]
10. Justen, A.; Cyganski, R. Decision-making by microscopic demand modeling: A case study. In Proceedings of the Transportation Decision Making: Issues, Tools, Models and Case Studies, Venice, Italy, 13–14 November 2008.
11. Varschen, C.; Wagner, P. Mikroskopische Modellierung der Personenverkehrsnachfrage auf Basis von Zeitverwendungstagebüchern. In Proceedings of the 7th Aachener Kolloqium "Mobilität und Stadt" (AMUS), Aachen, Germany, 7–8 September 2006; pp. 63–69.
12. Hertkorn, G.; Kracht, M. Analysis of large scale time use survey with respect to travel demand and regional aspects. In Proceedings of the International Association for Time Use Research Conference (IATUR), Lisbon, Portugal, 15–18 October 2002.
13. Hertkorn, G. Mikroskopische Modellierung von Zeitabhängiger Verkehrsnachfrage und von Verkehrsflußmustern, Dissertation. Available online: https://elib.dlr.de/21014/ (accessed on 11 August 2021).
14. von Schmidt, A.; Cyganski, R.; Krajzewicz, D. Generierung synthetischer Bevölkerungen für Verkehrsnachfragemodelle, ein Methodenvergleich am Beispiel von Berlin. In Proceedings of the Optimierung in Verkehr und Transport (HEUREKA'17), Stuttgart, Germany, 22–23 March 2017; pp. 193–210.
15. Statistisches Bundeamt, Mikrozensus 2015. Available online: https://www.destatis.de/DE/Methoden/Qualitaet/Qualitaetsberichte/Bevoelkerung/mikrozensus-2015.pdf;jsessionid=FE082318CCACFC78C8A0F48C95C0ED88.live712?__blob=publicationFile (accessed on 13 June 2021).
16. Senatsverwaltung für Bildung, Jugend und Familie, Zahlen Daten Fakten—Berufliche Schulen. 2018. Available online: https://www.berlin.de/sen/bildung/schule/bildungsstatistik/zahlen_daten_fakten_bs_2017_18.pdf (accessed on 8 September 2021).
17. Nexiga, Marktdaten—Demographie. Available online: https://www.nexiga.com/datensuche/marktdaten-demographie/ (accessed on 13 June 2021).
18. Nexiga, Marktdaten—Haushalte. Available online: https://www.nexiga.com/datensuche/marktdaten-haushalte/ (accessed on 13 June 2021).
19. Deutsche Rentenversicherung, Rentenatlas 2018. Available online: https://www.deutsche-rentenversicherung.de/DRV/DE/Experten/Zahlen-und-Fakten/Statistiken-und-Berichte/statistiken_und_berichte.html (accessed on 13 June 2021).
20. Amt für Statistik Berlin-Brandenburg, Statistisches Jahrbuch 2018. Available online: https://www.statistik-berlin-brandenburg.de/produkte/Jahrbuch/jb2018/JB_2018_BE.pdf, (accessed on 13 June 2021).
21. Henning, A.; Plohr, M.; Özdemir, D.; Hepting, M.; Keimel, H.; Sanok, S.; Sausen, R.; Pregger, T.; Seum, S.; Heinrichs, M.; et al. The DLR Transport and the Environment Project—Building competency for a sustainable mobility future. In Proceedings of the 4th Conference on Transport, Atmosphere and Climate, Bad Kohlgrub, Germany, 22–25 June 2015.