DRIVING PROFILE GENERATOR

Technical Report

by Benjamin Dietz, Klaus-Henning Ahlert, and Carsten Block

Last revised: June 28, 2010

Karlsruhe Institute of Technology

Faculty of Economics and Business Engineering

Institute of Information Systems and Management

Prof. Dr. rer. pol. Christof Weinhardt

Contents

List of Figures

List of Tables

1   Driving Profile Generator

The driving profile generator is a tool to generate an arbitrary number of car driving profiles such that the resulting profiles are in the same form as those reported in the German Mobility Panel (MOP) (BMVBS 2008), a large field study that reports car mobility for about 12.000 individuals in Germany. The profile generator described in this report is able to generate simulated driving patterns for four different groups of persons: employees, part time employees, retired people, and unemployed people. Each generated profile comprises all car movements of an individual within one week. For each of these trips several different pieces of information are provided, in particular the start of the trip, the duration of the trip, and the kilometers driven throughout the trip. To make the generated driving profiles follow patterns similar to those in the empirically observed driving profiles of the MOP, the panel's database has been statistically analyzed and the results are used as the basis for the profile generator. The following section describes in more detail how the MOP has been analyzed and how the major patterns have been identified. Thereafter, section 1.2 describes in more detail how the profile generator makes use of these statistical data to generate new driving profiles. Finally, section 1.3 evaluates the goodness of the generated driving profiles.

1.1   Analysis of the German Mobility Panel

There are three inputs for the profile generator that have been derived from the MOP. The first input is a database that determines the probability that there has been a trip on a certain day, for each day of the week for each group of people. For instance, the probability that an employee uses his or her car on a Monday has been determined.

The second input is the information about the probability of an occurrence of a certain *mobility-chain* on a certain day of the week. A mobility-chain contains all trips of one person on one single day. For example, such a mobility-chain could be "ToWork-ToHome". This mobility-chain represents the fact that the respective person did two trips on a particular day, i.e., one trip from home to work and afterwards a trip back home. In the following these components of a mobility-chain will be called *mobility-chain-components*. In the above example mobility-chain-component one is "ToWork" and mobility-chain-component two is "ToHome". The following mobility-chain-components are reported within the MOP:

1. ToWork
2. BusinessTrip
3. ToSchool
4. Shopping
5. Leisure
6. Service
7. ToHome
8. Mistake
9. ToHotel
10. To2ndHome

The probability for the occurrence of a certain mobility-chain on a certain day is calculated by dividing the number of occurrences of a certain mobility-chain by the number of occurrences of all mobility-chains on a particular day, with both information being reported in the mobility panel. In order to reduce complexity and to increase the number of mobility-chains within the period under observation, the different days of a week are grouped into the two different day types: *working day* (Monday to Friday) and *weekend day* (Saturday to Sunday). As can be seen in Figure 1 this grouping is a valid operation because driving behavior is very similar from Monday to Friday and from Saturday to Sunday, respectively.

Additionally, the mean, the standard deviation, the minimum, and the maximum of the starting times of the mobility-chains have been calculated. In order to determine solid values for these parameters weekdays and weekend days, again, have been merged as described above, hereby increasing the number of observations per mobility-chain. Furthermore, the original mobility panel database has been reduced to those mobility-chains that have the highest occurrence probability and account for 70% of all the observed mobility-chains during the respective observation period. Hence, rarely occurring mobility-chains are ignored and the main emphasis is put on modeling common driving behavior, for which many observations are available in the mobility panel's data. Similar to the grouping of weekdays, this data cleansing results in a larger minimum number of observations per mobility-chain. Overall, the cleaned panel data is still able to describe the most popular driving profiles, which account for 70% of all driving activity in all groups of persons and for working as well as for weekend days.

The concept of statistically analyzing the occurrence of mobility-chain probabilities
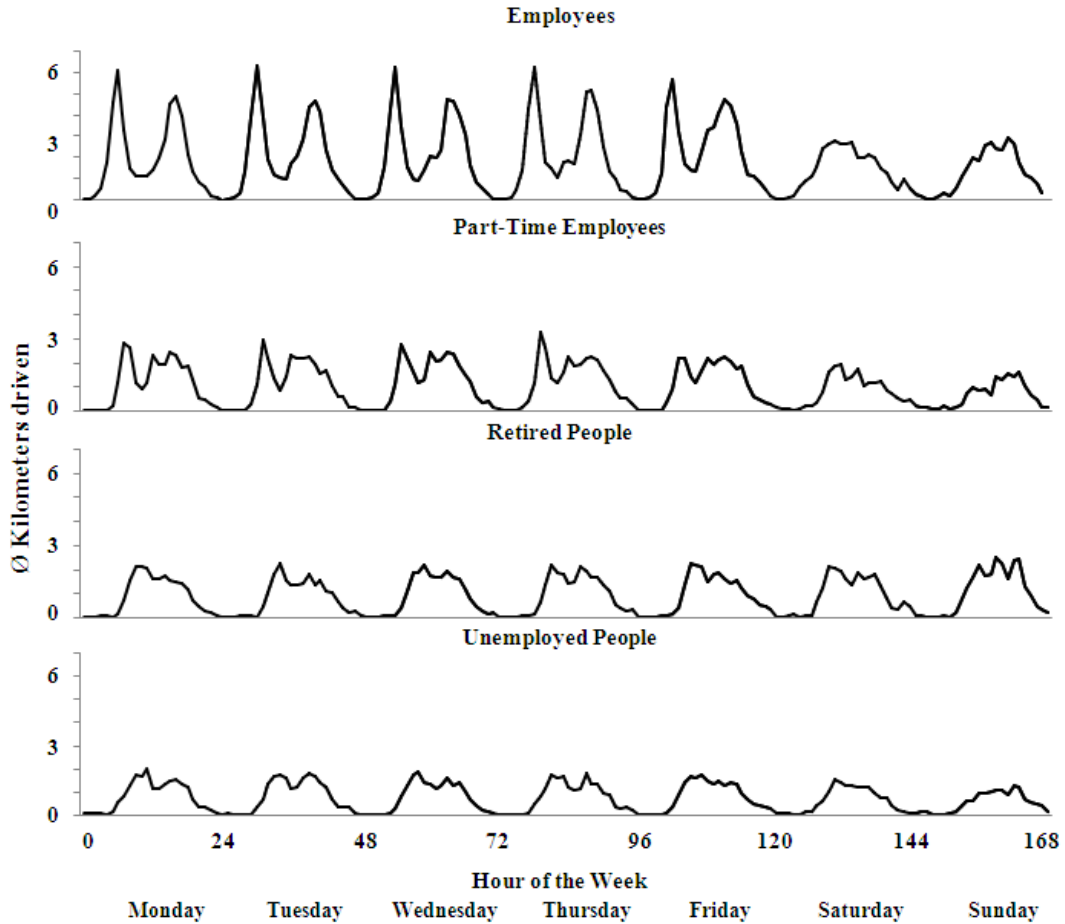
Figure 1: Average kilometers driven by different groups of people within each hour of the representative week

Source: Own analysis. 1000 most up-to-date driving profiles of each group of people from the MOP

and to then generate artificial driving profiles that possess the same diurnal driving probabilities requires mobility-chains to start and end on the same day. Hence, a mobility-chain that starts on a certain day also has to end on that same day before midnight. This is why driving profiles violating this constraint are excluded from the database. This reduces the number of driving profiles available for the analysis from 11,436 in the original MOP database to 7,674 (3,171 employees, 1,337 part-time employees, 2,077 retired people, and 1,089 unemployed people), i.e., 67% of the originally available driving profiles.

Figure 2 illustrates the effect of eliminating those driving profiles from the MOP database, where persons did not come home every day of the week. As can be seen, the overall average driving behavior of the employees group is different when comparing the original MOP panel data and the filtered data subset. However, only during weekends visible differences can be observed, the fit of the curves is still pretty good, which is also the case for the other person types. The correlation between the values (average kilome-

ters driven per each hour of the day) from the restricted database and the values from the unrestricted database is greater than 0.977 for each group of people.
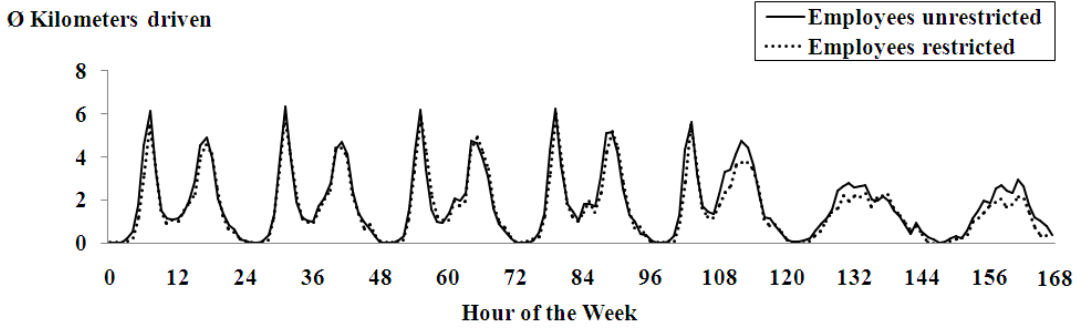

Figure 2: Unrestricted versus restricted MOP database. Average kilometers driven by employees within each hour of the representative week Restricted: Travels that pass 12AM are excluded

Source: Own analysis. Driving profiles from the MOP database.

The third type of input extracted from the MOP dataset and used as input for the profile generator is information about the occurrence of different mobility-chain-components within different mobility-chains. Each mobility-chain consists of several mobility-chain-components which all come attached with mobility-chain-component specific parameters. In particular, each mobility-chain-component contains information on the distance driven, the duration of the trip, average speed, and sojourn time, i.e., the resting time before the next mobility-chain-component starts.

For all mobility-chain-components specific parameters have been calculated: the mean value, the standard deviation, the minimum, and the maximum values. These specific parameters characterize the distributions of the mobiltiy-chain-components which can later be used artificially generating mobiltity-chain-components of a certain mobility-chain.

In the following the mobility-chain "ToWork-ToHome" will again be taken as an example. The two mobility-chain-components "ToWork" and "ToHome" are described through the above mentioned set of parameters. For instance, the distance that one particular person from the MOP has traveled from home to work could be recorded as being ten kilometers and five kilometers for another person respectively. Both persons then have the mobility-chain "ToWork-ToHome" reported in the MOP. If these people were the only ones in the overall panel's data set where this mobility-chain was recorded, the extracted statistical parameters for the two mobility chain components "ToWork" and "ToHome" would be $\mu = 7.5$, $\sigma = 3.54$, $min = 5$, and $max = 10$. These information are used later on to approximate the distribution of the kilometers driven in a certain mobility-chain-

component for the particular mobility-chain they occurred in.

For the duration, the speed, and the sojourn time the same descriptive statistics are calculated and stored. As mentioned earlier, in order to find significant parameters, it is important to have as many observations as possible. For these data the same data cleansing has been applied that was used before. The following Section describes in more detail how these three inputs from the MOP database can be used to generate artificial driving profiles.

1.2   Functionality of the Driving Profile Generator

The driving profile generator uses the inputs that have been described in the last section to step by step generate a one week driving profile like they are reported in the MOP. Due to the different driving behavior of different groups of people, all the inputs are calculated for each group of people (employees, part-time employees, retired people, and unemployed people). Hence, when generating a driving profile, first, the information for which group of people the driving profile shall be generated is needed. In the following the process of generating a driving profile for an employee will be explained. Afterwards, the starting weekday has to be selected - for example Monday. Thereafter, for each weekday the profile generator decides whether or not there is driving activity. This can be done using input one which contains the probability of a car usage by employees for each day of the week. Thus, if the probability of a trip on a Monday for an employee is 80%, the profile generator allows for trips on that day with 80% probability and does not allow for trips with a probability of 20%. The result could, for example, be that the person does not use the car only on Thursday.

As a next step, for all days except Thursday the profile generator has to generate the trips. At this point the concept of mobility-chains is used. For each of the days the profile generator selects one of the mobility-chains stored in input two (most popular driving profiles that account for 70% of all mobility-chains) according to their probabilities. For all working days the profile generator uses the same database of mobility-chains and their respective probabilities as it uses the same database for the days on the weekend. This is due to the grouping of days mentioned earlier. After this step, it has been determined at which day of the week the person is using the car and also for which mobility-chains.

Afterwards, the parameters of the mobility-chain-components (distance, duration, average speed, and sojourn time) and the starting times of the mobility-chains need to be

defined. Input two also contains the mean, the standard deviation, the minimum and the maximum of the starting time. The same specifications are provided in input three for the parameters of the mobility-chain-components. The idea is to find the distributions that can describe the characteristics of the respective parameter and then use these distributions to generate random values for the parameters. Figure 3 shows a histogram of the distance driven by employees on the way to work. This histogram is specific to the case of the mobility-chain "ToWork-ToHome" and the days from Monday to Friday. For a different mobility-chain, group of people, or time frame, the histogram will look different.
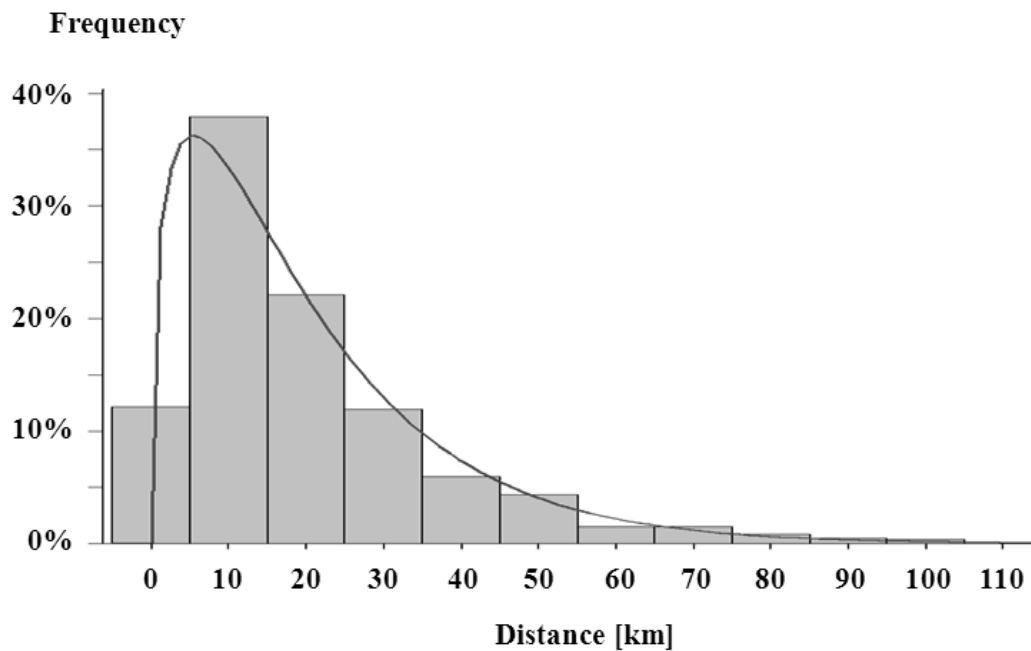

Figure 3: Histogram of the distance of the way from home to work within the mobility-chain "ToWork-ToHome" on a working day for employees

Source: Own analysis. Driving profiles from the MOP database.

After a lot of curve fitting tests it turned out that none of the common distributions could fit the data significantly according to tests like the Kolmogorov-Smirnov test, the Cramer-von Mise test, or the Anderson-Darling test. However, as can be seen in figure 3, the gamma distribution, illustrated by the line above the bars of the histogram, provides a pretty good fit (as an approximation) for that particular case. For other parameters like the duration, the speed and the sojourn time, the gamma distribution also provides a rather good fit as well, although being not perfectly accurate. Figure 4 illustrates the fit of the gamma distribution for the duration within the same situation as before.

The situation is rather complicated since there are a lot of different mobility-chains with many different components, and each component consisting of several parameters.
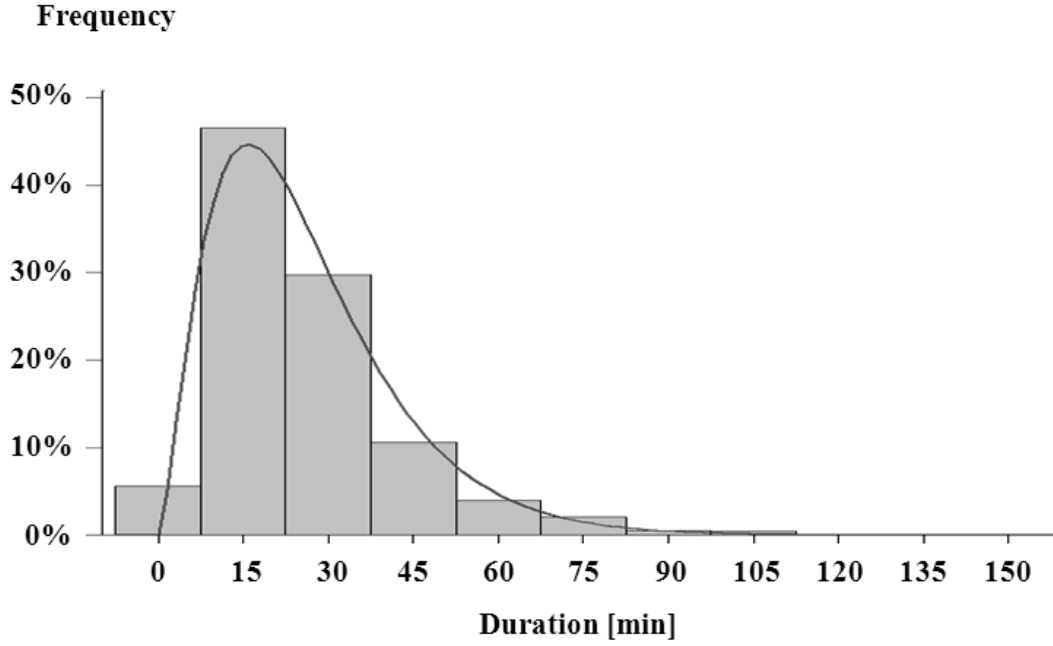
Figure 4: Histogram of the duration of the way from home to work within the mobility-chain "ToWork-ToHome" on a working day for employees

Source: Own analysis. Driving profiles from the MOP database.

For each of these parameters the distribution has to be found that fits the best. The ideal case would be to have a curve fitting test, that is done for each parameter independently. However, such an approach would increase the complexity of the profile generator significantly. In order to keep the profile generator as simple as possible only the gamma distribution has been used as an approximation for all the parameters of the mobility-chain-components and the starting time of the mobility-chains. The probability density function of the gamma distribution looks as follows:

$$f(x) = \begin{cases} \frac{e^{-\lambda x}\lambda^k x^{k-1}}{\int_0^\infty t^{k-1}e^{-t}\,dt} & : x > 0 \\ 0 & : \text{otherwise} \end{cases} \tag{1.1}$$

The parameters λ and k can be calculated from the mean (μ) and the standard deviation (σ) using the following equations:

$$\lambda = \frac{\mu}{\sigma^2} \tag{1.2}$$

$$k = \frac{\mu^2}{\sigma^2} \tag{1.3}$$

λ^{-1} is also referred to as the scale parameter whereas k stands for the shape parameter. Using the equations 1.2 and 1.3, the parameters of the gamma distribution can be calcu-

lated with the mean and the standard deviation from input one and two for the respective parameter. Having the distribution of a certain parameter specified, random values can be generated using this distribution. Taking the example of the mobility-chain "ToWork-ToHome" again, the distance driven by employees on their way to work on a working day follows the gamma distribution with a certain shape parameter and a certain scale parameter. For a new employee driving profile this distribution can be used to generate a random number for the distance driven facing exactly the same situation (same mobility-chain and same trip). Hence, when a large number of driving profiles for employees is generated, a histogram of the distance driven during a certain trip by these employees will follow the gamma distribution that has been used to generate the values for these parameters.

Therefore, the mean and the standard deviation provided in input one and two are used to calculate the shape parameter and the scale parameter of the gamma distributions for the respective parameters. The resulting distributions are then used to generate random numbers for the parameters of new driving profiles. The inputs two and three also include the minimum and the maximum value of the observed parameters. These values are used to avoid outliers generated from the distribution that do not exist in reality. Hence, every time the value for a parameter exceeds the limits (lower than the minimum or higher than the maximum), a new random value is generated. This is done until the calculated value lies in between the limits of the parameter. As a result, this mechanism produces only realistic values. Figure 5 illustrates how this mechanism would restrict the gamma distribution for the scenario that has been shown already in figure 4.

Since the duration, the speed, and the distance are not independent of each other, but rather each one is determined by the two others, not all three parameter values are generated using the gamma distribution. The speed and the distance are generated using the gamma distribution whereas the duration is then calculated through the other two parameters. If all three parameters would be generated independently, this could lead to inconsistencies.

Using the proceeding described above to generate all the parameter values for the components of the mobility-chains and the starting times of the mobility-chains on each day of the week, provides the whole one week driving profile. If just this would be the functionality of the profile generator, it would generate driving profiles that produce a similar average driving behavior like the driving profiles from the MOP. However, it is also important, that each single driving profile seen individually also "makes sense".
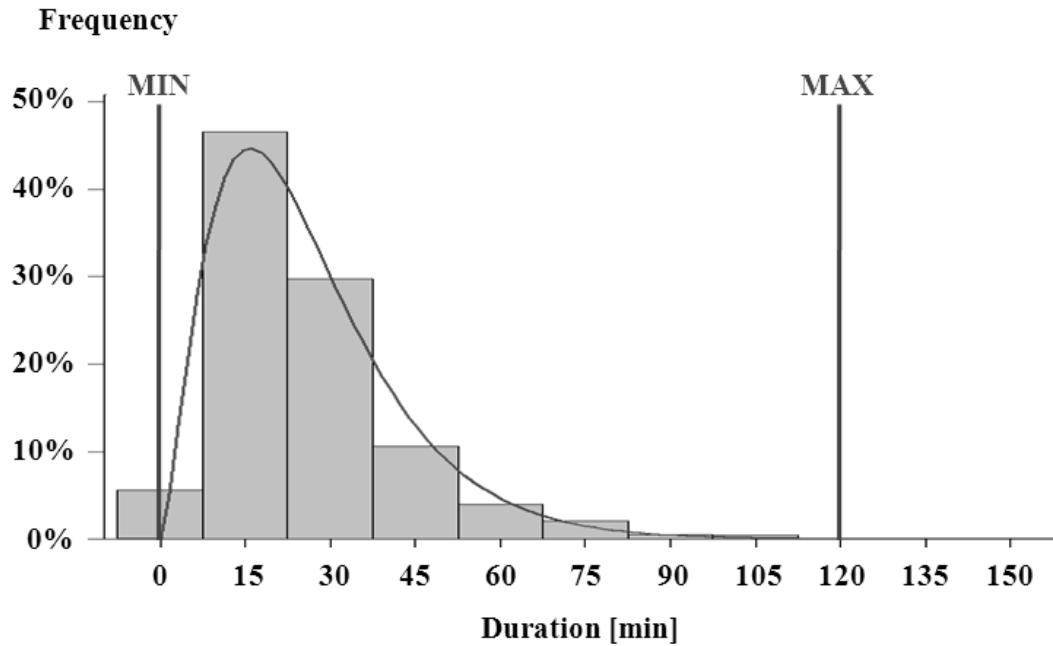
Figure 5: Histogram of the duration of the way from home to work within the mobility-chain "ToWork-ToHome" on a working day for employees including the limitations of the parameter

Source: Own analysis. Driving profiles from the MOP database.

The functionality described so far would allow for a situation in which on Monday a person drives 10 kilometers from home to work and on Tuesday the same person drives 100 kilometers for the same route. Hence, the following rules that take account of such dependencies within a single profile, have been developed:

1. The distance between home and work has to be similar throughout the whole week if the person directly drives from home to work or vice versa. Hence, the first time the person directly drives from home to work or from work to home, the distance between home and work is fixed to a certain value. The next time the distance has to be specified for a trip between home and work, the distance that has been fixed earlier, is taken. In order to allow for some variation, a factor between 0.8 and 1.2 is multiplied with this distance. The factor is generated as a random number from a normal distribution with a mean of one and a standard deviation of $\frac{1}{15}$.

2. As already stated in rule 1, the distance between home and work has to be similar throughout the whole driving profile. In case of an indirect way between home and work, certain limits for the distance between home and work can be calculated. An indirect way means that a person, for example, stops at the supermarket on the

way from home to work. It is possible that such a stop is just on the way and no additional kilometers have to be driven. However, the supermarket can also be situated in the exact opposite direction of the working place. Hence, on one day the limits for the distance between home and work are calculated so that on the next day these limits can be considered. The upper bound of the distance between home and work is the sum of all the distances of the ways between home and work and vice versa whereas the lower bound is the distance of the last trip minus the sum of the distances of all previous trips. Of course, the lower bound can never be less than zero. The next time, a direct trip or an indirect trip between home and work has to be determined, these limits have to be considered.

3. For a roundtrip, i.e., a trip from home to home or from work to work including other stations in between, the distance of the last trip has to lie within certain limits. The distance of the last way has to be larger than the distance of the first way minus the sum of the distances of all other previous ways. This lower bound can be interpreted as a situation in which the person first drove away from the starting point and then always drove towards the starting point again. Of course, the minimum of the distance of the last way is zero. Hence, if the calculated minimum is smaller than zero, it is set to zero. The upper bound of the distance of the last way is the sum of all previous ways. This represents the case that the person always drove further away from the starting point and the last trip goes all the way back to the starting point again. As before, the gamma distribution is used to generate a value for the distance, but the result has to lie in between the calculated limits, otherwise another random value is generated with the gamma distribution until it lies within the limits. This rule is only applied to roundtrips from and to work and from and to home, since those are the only unique places that can be identified within the driving profiles for employees, part-time employees, retired people and unemployed people.

4. Since mobility-chains are developed day by day and each day has only one mobility-chain, the mobility chains have to end before midnight, so that they cannot influence the driving behavior of the next day. Otherwise mistakes could occur due to an overlap between different mobilitiy-chains. Furthermore, the input of the profile generator are driving profiles that only contain mobility-chains which end before midnight on each day. Due to these two reasons, also the generated mobility-chains

have to end before midnight. In case a mobility-chain does not end before midnight, all the components are calculated again.

5. Rule 1,2, and 3 generate limitations for the distance between home and work. It can be the case that these limitations are generated at the beginning of the week and do not fit with the a mobility-chain of the following days. It is, for instance, possible that on a Monday the distance between home and work is set to 10 kilometers. If for Tuesday a mobility-chain has been selected that contains a direct trip from home to work with a minimum distance of two and a maximum distance of four kilometers, the 10 kilometers from Monday are not feasible for that particular mobility-chain-component on Tuesday. Hence, the combination of mobility-chains that have been generated for each day, do not fit with each other. If such a situation occurs, the whole week is started again and new mobility-chains are selected from the database for each day of the week.

6. In order to create a person specific driving profile for one week, another effect is considered. Since the mobility-chains are selected according to their probability for each day of the week, it is likely that the sojourn time at work differs between different days of the week. However, there are people who tend to work longer hours than the average person and there are also people who tend to work less hours than the average person. Therefore, another factor is included that is multiplied with the sojourn time at work for each day of the week. This factor is calculated in the same way as the variation factor in rule 1. If a factor of 0.9 is generated at the beginning of the week, this factor is used to reduce the sojourn time at work of each day by 10% for that particular person.

Using these six rules for the generation of one week driving profiles, each individual driving profile makes sense in terms that very unrealistic situations are avoided.

1.3   Evaluation of the Driving Profile Generator

Every model has to abstract from reality. The following examples show where the profile generator abstracts from reality and, therefore, is not 100% accurate:

- The mobility-chains on working days are grouped as well as the mobility-chains on the days of weekend although the driving behavior on these days may be slightly

different for each day. It would be more precise to analyze the driving behavior day by day. However, by grouping these days, the number of observations for each mobility-chain increases. This improves the approximation of the distribution function of the mobility-chain-component parameters.

- Driving profiles that are used for the profile generator are restricted to those people who come home every day. This leads to inaccuracies in the overall driving behavior of all people (see figure 2). In order to avoid this effect another approach than the mobility-chain approach would need to be developed.

- The curve fitting, exemplary shown in figure 4 and 5 in 1.2, is not perfectly accurate due to the assumption of the gamma distribution for all parameter values. Ideally, each parameter should be estimated with the individual distribution function that fits best. However, the large number of parameters makes this approach much more complex.

- The number of rules that have been developed are limited to six rules. In reality there are probably many more rules and dependencies that are not considered within the profile generator. However, the more rules are applied, the more predictable are the driving profiles and the more complex the model becomes. If all the information about the driving profiles from the MOP would be considered, only driving profiles that occur already in the MOP could be generated. This would mean that the same driving profiles are used over and over again.

Figure 6 shows a comparison between the average driving behavior of employees from the restricted database (people coming home before midnight) and the average driving behavior of employees generated with the driving profile generator. It can be seen that the profile generator does not generate the exact driving behavior as can be found in the MOP, due to the simplifications mentioned above. However, it can also be seen that the general patterns are captured by the profile generator.

In general, the profile generator tends to underestimate the kilometers driven. This effect can be observed when comparing the average kilometers driven during one week. Table 1 provides a comparison between the average kilometers driven per week by the different groups of people. Therefore, when using the driving profile generator to generate the driving behavior of people, this effect has to be taken into consideration. However,
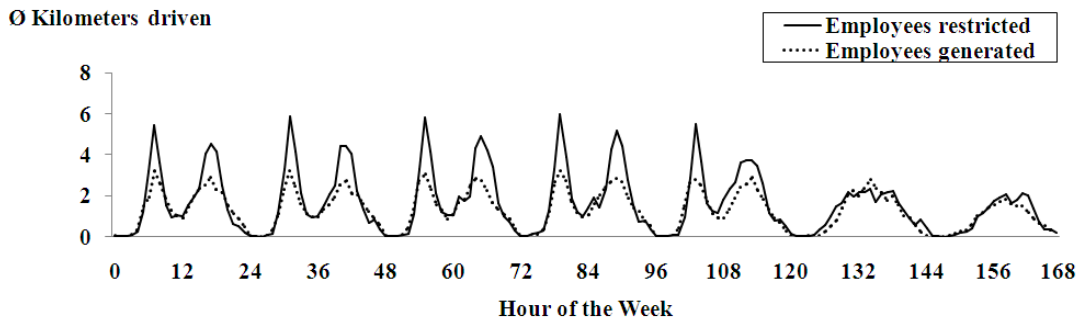
Figure 6: 1,000 generated driving profiles versus 1,000 driving profiles from the restricted MOP database. Average kilometers driven by employees within each hour of the representative week

Source: Own analysis. Driving profiles from the MOP database and the driving profile generator.

Table 1: MOP versus Driving Profile Generator. Average Kilometers driven per Week

	Unrestricted Database	Restricted Database	Driving Profile Generator
Employees	302 km	274 km	215 km
Part-Time Employees	173 km	159 km	136 km
Retired People	153 km	141 km	130 km
Unemployed People	123 km	114 km	86 km

the driving profile generator has primarily been developed to generate driving profiles of people using an electric vehicle (EV). In general, it is likely that in the near future people who use an EV instead of an ICEV, drive less kilometers on average, since the specifications of today's EVs are more suitable for those people than for people who drive long distances.

Although abstracting from reality at some point, the profile generator produces driving profiles that are much alike the real driving profiles from the MOP. In general, profiles from the profile generator can be used for the driving behavior of people using any kind of vehicle. However, the results shown in table 1 illustrate that the driving profiles reflect the driving behavior of people who on average drive shorter distances than today's car owners.

Initially, the profile generator has been developed as a web application at the Institute of Information Systems and Management (IISM), Karlsruhe Institute of Technology (KIT), Germany. It can be accessed online at `http://ibwmarkets.iw.uni-karlsruhe.de/ps`. Alternatively a java-based stand-alone version of the generator is available, which allows for generating large numbers of mobility profiles as xml or csv files. For further

information on availability of the software please contact the "Telecommunications & Energy" research group at IISM (`http://www.im.uni-karlsruhe.de`).

References

BMVBS (2008). German Mobility Panel (Deutsches Mobilitätspanel), Panelauswertung 2007. Deutsches Bundesministerium für Verkehr, Bau und Stadtentwicklung. `http://mobilitaetspanel.ifv.uni-karlsruhe.de`. Last visit: 03/17/2010.