TECHNISCHE UNIVERSITÄT
KAISERSLAUTERN

Fraunhofer
ITWM

Automatic Usage Modeling for Automotive Applications

Vom Fachbereich Mathematik
der Technischen Universität Kaiserslautern
zur Verleihung des akademischen Grades

Doktor der Naturwissenschaften
(Doctor rerum naturalium, Dr. rer. nat.)

genehmigte
Dissertation

von
Christine Biedinger

Gutachter:
Prof. Dr. J. Franke, Technische Universität Kaiserslautern
Prof. Dr. G. Lindgren, Lund University, Schweden

Ort und Datum der Disputation:
Kaiserslautern, 11.Oktober 2018

D 386

Acknowledgments

My thanks go to all those who supported me in the completion of this work, either in professional or personal way.

First of all I would like to thank my supervisor, Prof. Dr. Jürgen Franke, for his kind advise and guidance. This thesis was developed during my time as Ph.D. student at the Fraunhofer Institute for Industrial Mathematics in Kaiserslautern in the department *Mathematical Methods in Dynamics and Durability*. I would like to thank Dr. Klaus Dreßler, the head of the department, for the opportunity to work on this interesting topic and the financial support. Furthermore, I would like to thank my colleagues for the pleasant atmosphere. In particular, the VMC team shall be mentioned here. I would also like to thank Dr. Michael Speckert and Thorsten Weyh for the valuable discussions and productive cooperation in the application of the developed usage models. Special thanks go to Dr. Sascha Feth who was my advisor at ITWM not only for my Ph.D., but also during my whole studies. Thank you for your continuous support and encouragement and the fruitful discussions.

Abstract

Many loads acting on a vehicle depend on the condition and quality of roads traveled as well as on the driving style of the motorist. Thus, during vehicle development, good knowledge on these further operations conditions is advantageous. For that purpose, usage models for different kinds of vehicles are considered. Based on these mathematical descriptions, representative routes for multiple user types can be simulated in a predefined geographical region. The obtained individual driving schedules consist of coordinates of starting and target points and can thus be routed on the true road network. Additionally, different factors, like the topography, can be evaluated along the track.
Available statistics resulting from travel survey are integrated to guarantee reasonable trip length. Population figures are used to estimate the number of vehicles in contained administrative units. The creation of thousands of those geo-referenced trips then allows the determination of realistic measures of the durability loads.
Private as well as commercial use of vehicles is modeled. For the former, commuters are modeled as the main user group conducting daily drives to work and additional leisure time a shopping trip during workweek. For the latter, taxis as example for users of passenger cars are considered. The model of light-duty commercial vehicles is split into two types of driving patterns, stars and tours, and in the common traffic classes of long-distance, local and city traffic.
Algorithms to simulate reasonable target points based on geographical and statistical data are presented in detail. Examples for the evaluation of routes based on topographical factors and speed profiles comparing the influence of the driving style are included.

Contents

1. Introduction

1.1. Motivation and problem description

The usage of vehicles is very variated. There exists a large number of different types, special equipments and diverse fields of application. Thus, during vehicle development specific knowledge on the future operating conditions is needed. The forces acting on a vehicle have to be estimated realistically to enable design decisions fulfilling the required demands. These forces strongly depend on the roads traveled as well as on the driving style of the motorist.
The common approach conducting a measurement campaign in the region for which the vehicles are constructed has multiple disadvantages. It is expensive, requiring a lot of money and time. Additionally, it is not directly clear if the roads passed are representative for the considered usage type and how the obtained measurements have to be combined to achieve reliable estimates for the expected damage values. Furthermore, at an early stage of development, there might be no prototype available which can be used for the campaign.
In an alternative situation, different regions shall be compared. This is then still more expensive. If emerging markets shall be considered, reliable information on the vehicle usage is also expected to be missing.
In this work we solve these problems by gaining the required knowledge virtually. We develop usage models for different kinds of vehicles and apply mathematical methods to simulate the routes of several thousands typical drivers. We integrate different kinds of statistics in order to create representative routes and use geographical data allowing the evaluation of attributes depending on the routes traveled.
The advantages of this approach are versatile. First of all, it is not restricted to a specific region. Once a usage model is at hand, it only has to be adapted slightly to the new area of interest and the simulation can be restarted directly. Furthermore, the influence of different factors can be compared. In the usage model for commuters for instance, trips to shopping locations are included. Here, the difference between the application of a distance distribution resulting from some travel survey and the selection of the nearest shopping facility could be investigated. Additionally, also parameters of the vehicle can be varied for the routing between the computed origin-destination pairs. Shortest and fastest connections can be compared as well. For the measurement campaign, one has to decide before the start which forces to measure. In our approach, a subsequent supplementary

evaluation of the roads covered is always possible. The results obtained from the simulation of usage models can be used in combination with already conducted campaigns and enhance the results. They provide the missing settings for the computation of pseudo-damage values in the usage simulation.

We want to clarify that the goal of usage modeling lies not in the creation of individual really existing travel schedules. We rather want to reproduce the vehicle population of interest as a whole. Only the summarized outcomes for a sample size large enough are significant.

1.2. Outline of this work

The overall problem of creating representative routes automatically is split into several subproblems in the following. We start with the development of different usage models. First, we concentrate on light-duty commercial vehicles. We demonstrate the complete process from preparing required inputs, simulating and evaluating representative routes. At last we show how they can be combined with measurement campaigns. This chapter extends the methods already presented in [22] and also [60].
Afterwards, we consider passenger cars. We repeat the model of commuters describing the common usage of private passenger cars introduced in [21]. Additionally, we sketch the preparation and integration of available traffic surveys. We finish the chapter with a usage model for taxis as example of users of commercial passenger cars covering large distances during a week conducting various types of trips.
In chapter 4 we consider different kinds of data required for the simulation. We start with geographical information forming the basis for the calculation of origin-destination pairs. We distinguish between point and areal data and describe mathematical methods to select suitable coordinates. Afterwards, we show that, at least for Germany, population figures can be applied to estimate vehicle distributions. We also demonstrate that the population counts available in map data used are of sufficient quality to replace official population statistics that otherwise had to be integrated in the database. We finish the chapter with a look at distance distributions. On basis of the travel survey "Mobilität in Deutschland 2008" (Mobility in Germany 2008)[6] we describe the analysis and integration of such data.
Chapter 5 treats different aspects concerning the implementation of the algorithm. We start with the mathematical method for the choice of home locations of vehicles. We introduce different techniques for the modeling of population distributions found in literature. We then create a method exploiting the characteristics of the available geographical data. Next, we try to find influencing parameters for detour factors which indicate the dissimilarity of linear and driven distances. We also sketch how these can be used to integrate the distribution of driven distances

from traffic surveys in our techniques based on linear distances. At last, we consider the situation that our algorithm is not able to create a feasible trip chain. This could happen if a distance is prescribed in which no adequate target point is located.
In chapter 6 we demonstrate the algorithm of creating representative routes with the example of a commuter living in Kaiserslautern. We depict how his driving schedule for a complete workweek is assembled. Like in chapter 2.6, topographical factors are analyzed for the obtained routes. In the second part of that chapter we consider a taxi driving around Kaiserslautern. Here, two speed profiles for a single trip are created, one for an aggressive and one of a careful driver. It is sketched that the driving style of the motorist has a remarkable influence on the forces acting on the vehicle driving on the same roads.

For this work we use VMC®, a software developed at ITWM, for the routing and analysis of covered roads as well as for the simulation of the speed profiles. The methods applied are taken as provided, applying standard settings. We also extract the geographical data introduced in chapter 4.1 and the population counts employed in chapter 5.1 from the VMC® database.

1.3. Related work

Some of the problems to be solved for the simulation of usage models also have been considered by different people. The list of sources does not claim to be comprehensive, only an extract shall be reviewed here.
Most literature can be found for the commuter model. Daily travel patterns are analyzed and simulated for more than forty years. The regular conduction of traffic surveys shows that the interest in current data still legitimates the costs and effort required. However, usually no concrete locations represented by their geographical coordinates are treated, like they are needed in our simulation. Mostly, households and the persons belonging to them are created like in [20], [23] or [45]. The spatial classification is at the most performed on basis of zones therein. The overall population is simulated such that socio-economic attributes are reflected well. Beckman, Baggerly and McKay obtain geographical coordinates in [19], but they follow a different approach. They extract subsamples from available original census data. Hay et al. [34] instead consider the locations of people like we do, even they require them for public health applications. Trip chains are not considered there.
They are the central theme for Bhat et al. [20] and Lovelace et al. [45] again. In the former source time is used as critical attribute. In the latter shortest routes are assumed but they are just computed between zones, not specific points. This does not match our needs for reliable values for covered distances. Susilo and

Kitamura [57] work with both, trip chains and driven distances, but on one hand they restrict their examinations on the cities of Karlsruhe and Halle, Germany. On the other hand they introduce the concept of *action spaces*, "a set of places where an individual visits to carry out activities" [57].
Literature dealing with the creation of population densities shall be skipped here since a comparison of different methods is given in chapter 5.1. Summarizing, there is active research in the same direction like the usage models, but the focus always lies on different aspects lacking of some attributes required for our purposes. Especially the geographic component seems to be mostly neglected.
A second concept that has gained a lot of interest is the one of detour or circuity factors. Phibbs and Luft [48] compared the linear distance with travel time and stated that with decreasing distance the time to travel one mile grows up. Ballou et al. [18] replaced the time aspect by the distance covered. Giacomin and Levinson [32] consider both measures. Concerning the factors based on distances only, two different methods to compute the average are applied. Ballou et al. compute the detour factors for inter-city distances first and then calculate the mean of all values. Giacomin and Levinson divide the sum of network distances by the sum of Euclidean distances. For our purposes we work with the concept of Ballou et al.. Levinson and El-Geneidy [43] perform detailed comparison of the difference in circuity between randomly sampled points and true work-home pairs and state that the second average is significantly lower than the first. However, again only metropolitan areas are considered. This is not sufficient for our calculation where people are simulated in rural districts also. Hence, [18] is used as a basis but the applied factors should be adapted to our purposes.

2. Usage model for light-duty commercial vehicles

In this chapter, we introduce our approach in a rather simple setting by looking at a specific example, the simulation of light-duty commercial vehicles. It extends the rather tersely described concepts given in [22].

2.1. Introduction

Even if usage modeling is concentrated on light-duty commercial vehicles, such as parcel services, craftsmen or long-distance services, the daily utilization of motorcars is strongly varying. The differences mainly lie in the commercial sector. Delivery services for example are often faced with time restrictions because goods have to be delivered in a certain time slot. Diverse customers are waiting in line, not accepting delays. Craftsmen on the other hand only have to meet their working hours. They are expected to visit fewer customers per day and often have a less strict time schedule. The typical trips of these motorists during their shift show different driving patterns. The goal in vehicle development is to guarantee a desired target mileage for all motorcars. In order to reach this, reliable estimates on damage values are requested. Those are influenced by the forces acting on single components of a vehicle. Depending on the usage group and the derived characteristics, these forces are expected to differ. The following chapter describes the process how the required damage values can be estimated. Therefore representative routes reflecting ordinary vehicle life depending on industrial sector and motorist are computed. We concentrate on the modeling and simulation of the vehicle usage for light-duty commercial vehicles and show how these results can be combined with measurement campaigns.

2.2. Characterization of model

The goal of modeling vehicle usage is the generation of artificial data representing the typical use of, in our case, light-duty commercial vehicles. Each individual observation which we want to generate represents a route R of a vehicle over

several days. We start with a network of roads of various categories like highway, country road etc. which mathematically correspond to a planar graph G with marks on the edges. These marks not only signify the road category, but also other information, e.g. if the road lies in a residential or in a commercial area.
The route

$$R = R(GI, MP, IV) \tag{2.1}$$

which we want to generate or, more precisely, its distribution, depends on three sets of parameters, where GI stands for geographical and user identification, MP for model parameters and IV for, usually random, input variables,

$$GI = (region, index) \tag{2.2}$$

$$MP = (classlimits, classcharacteristics) \tag{2.3}$$

$$IV = (U, P, D, n, B, S) \tag{2.4}$$

which we distinguish due to their different roles:
GI consists of control parameters which we choose to suit the goal of the particular simulation, e.g. to simulate routes in a particular region as subset of the whole road network graph G.
MP contains parameters which are fixed somewhat arbitrarily. We partition routes with regard to the maximal distance traveled into three classes (long, medium, short) where there is no natural or established rule for choosing the separating points. Table 2.1 contains examples for such class limits depending on the shape of the route. E.g. for a “tour” defined later, the "medium distance" class consists of routes with linear distance between 20km and 50km between the start location of the route and approached stopovers. MP also contains parameters of the distribution of inputs from IV. Due to lack of appropriate data, these parameters are mainly set to plausible but yet arbitrary values but, preferably, have to be replaced by estimated ones in the future once appropriate data are available.
Finally, IV contains the random input variables into the system generating the route. They consist of various qualitative random variables like shape of the route or distance class, discrete numerical random variables like number of stopovers and continuous random variables like the coordinates of the home base of the vehicle and of the stopovers on the route. The joint distribution of those inputs has to be estimated which is not trivial as they depend on each other. We later discuss which data would be needed ideally and which data is currently available and how they may be used to approximate the distributions of the inputs as the basis for the simulation.

First, we provide some more details about the parameters and inputs. The introduced values have the following interpretation:

- ***R***: A route R consists of a list of significant points given by their latitudinal and longitudinal coordinates. The driving schedule begins with the origin or "home" location where the vehicle is parked usually at start of work. Subsequently, stopovers visited during a shift are listed. These might be job-sites for craftsmen or the positions of customers for delivery services. Those points are approached according to their order, thus the last destination is again the parking position. Of course, the stopovers need not to be nonrecurring. Craftsmen for example often return to the office between jobs, delivery services carrying food always need to retrieve the next offers.
- ***region***: The geographical unit in which the vehicles drive around. This might be a country, a federal state, a town or even a artificially created combination. The region needs to be represented by a "multipolygon" consisting of a description of its boundary. See section 4.1.6 for detailed information on this datatype.
- ***index***: Consecutive number for identification of single vehicle in bulk of simulations. Combination of *region*, *index* and user U forms a unique identifier.
- ***classlimits***: Fixed bounds for distance classes used in simulation. For all cases of driving pattern P suitable values are required. This variable is directly connected to *classcharacteristics* providing further information on the classes. An example for the split in three classes is given in the first row of table 2.1.
- ***classcharacteristics***: Values further describing the distance classes contained in simulation. These include the expected distribution of trip lengths regarding *classlimits* and corresponding pattern P. Additionally, the distribution of the number of customers visited on one trip is reported. For each class given in *classlimits*, all associated values have to be provided. Table 2.1 shows an example for a complete configuration of the model parameters MP.
- ***U***: User of vehicle. Depending on the simulated vehicle type or industrial sector, the usage differs. For light-duty commercial vehicles, "craftsmen" and "delivery services" are common values of this variable.
- ***P***: Driving pattern. The order of points visited during one or multiple trips shows a specific pattern. For light-duty commercial vehicles "stars" and "tours" and a combination of both are typical shapes. "Stars" are characterized by alternation between a central destination like the home of the vehicle and varying further positions. An ideal example for this driving pattern is that of a craftsmen who visits several customers but goes back to his office after each job. "Tours"' are usually performed by delivery services visiting multiple customers one after another before they return to their base. Mixtures of both patterns contain some locations that are visited repeatedly

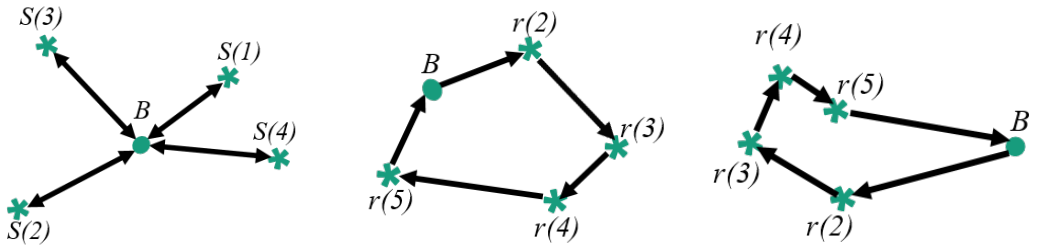
without returning home in between but driving to further places. Those can be cut smartly into pieces and be treated as combination of the base patterns. The general form of "stars" and "tours" is sketched in figure 2.1. It includes the special case that the depot is located outside a town, usually easy to reach by road, and the customers live in the city center. The pattern is a bit degenerated then but still tour-shaped. Examples for the base types projected on the map are given in figure 2.2. The actual driving pattern $P = i$ is depending on the chosen user $U = j$.

- $\boldsymbol{D}$: Actual distance class. The probability for a distance class $D = k$ depends on chosen user and pattern, $Pr\,(D = k\,|U = i, P = j\,)$. At the moment, the frequencies are prescribed in $classfrequencies$ and are reflected exactly when simulating a population of vehicles. A treatment as frequency distribution allowing deviations can be integrated.

- $\boldsymbol{n}$: Number of stopovers. This variable determines the count of customers who are visited on the specific simulated trip. Up to now, this number is taken as the constant value set in $classcharacteristics$. In the future, more variation shall be added. One possible data source might be the survey "Kraftfahrzeugverkehr in Deutschland" [26]. The public use file allows to count the number of routes conducted per day by a single vehicle, see section 2.3.3. The distribution of the number of stopovers can then be estimated with respect to various constraints like day of week or industrial sector. n always depends on these influencing factors, $n = n(U, D)$.

- $\boldsymbol{B}$: Base location of vehicle given as geographical coordinates, $B \in S_2$. Usually, a light-duty commercial vehicle has a common parking position where it is returned to after end of shift. This location often lies near the companies site. The specific place is chosen in a commercial, industrial or retail area as well as in a residential one if the vehicle belongs to a small family company for instance. It can also be chosen from a list of businesses registered in the considered $region$. More information on available data sources and the corresponding data types is given in chapter 4.1.

- $\boldsymbol{S}$: List of stopovers. $S \in \mathbb{R}^{n\times 2}$ gives the coordinates of all simulated destinations except B. Depending on P, the list is reordered after creation of all positions. For $P =$"star" each customer is visited individually resulting in n single trips to one stopover each and returning home afterwards. If $P =$"tour", all places have to be visited one after another before returning home. This creates $n + 1$ tracks forming one single trip starting at B.



(a) Star (b) Tour (c) Tour with center outside

Figure 2.1.: Outline of the two patterns, own illustration from [22].

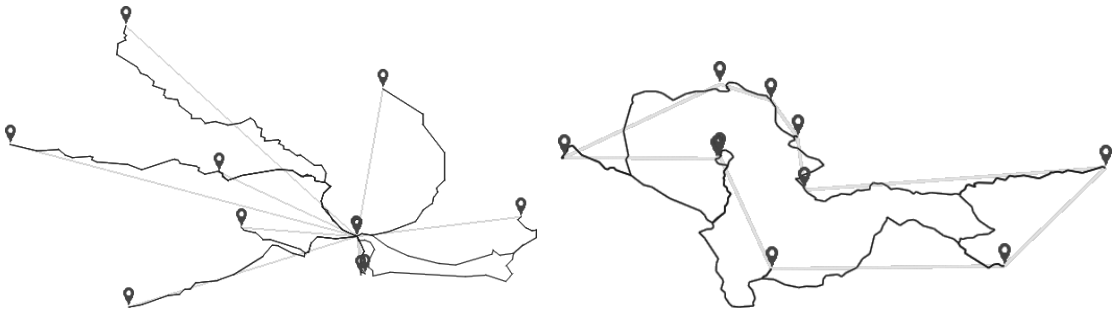


Figure 2.2.: Examples of base patterns given in [22]. Simulated points are connected linearly by gray lines, routes on the road network are painted black.

2.3. Determination of values for MP

2.3.1. Background of distance classes and patterns

In the usage model described above driving patterns play an important role. Usually, traffic classification is only based on driven distances but does not include such a concept. In this section, the usefulness of the new approach is illustrated. The usage of light-duty commercial vehicles typically is divided into the three categories city, local and long-distance traffic. This legal partition into different distance classes was disestablished in Germany more than 40 years ago, but the terms are often still in use. When having a closer look at typical trips, some drawbacks of this classification for usage modeling get obvious since it is only based on the operation range of the vehicle. Like it is shown in figure 2.3, people working in various industrial sectors perform journeys in more than one distance category regularly. However, in all three categories we can determine similar driving patterns for vehicles with the same trip purpose, just differing in scaling.
A craftsman for instance has a catchment area depending on the structure of urban development in his place of residence. If he is living in a large city, he might find enough customers in a short distance. If he is living in the countryside, he

presumably has to accept orders in a larger region. The same holds for delivery services. Considering postal services, one can distinguish between parcel and letter delivery having different operating radii. Postmen visit nearly every house, whereas parcel carriers stop at addresses further apart. Thus, those can serve a larger area. In both examples, the driven distance and the number of customers visited varies, but the basal pattern stays the same.

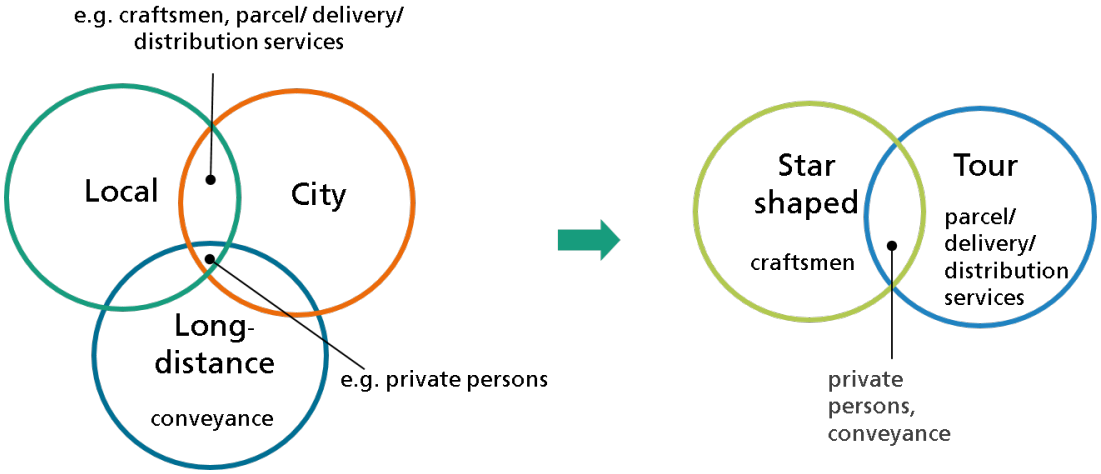


Figure 2.3.: Classical and new division of traffic, own illustration from [22]

If the traditional subdivision into three traffic types is necessary in the end, a translation between both partitions is possible. First of all, the model parameters MP can incorporate the old concept by setting three entries for *classlimits* for each pattern. Additionally, *classcharacteristics* allows constant frequencies or differentiates between the patterns. Hence, it is possible to characterize one traffic class as a combination of patterns with different distance bounds. The weighting of all components allows the incorporation of the influence of the industrial sector. Returning to our standard example, city traffic for the building industry in contrast to postal or delivery services for instance can be composed of the same distance classes but can have different proportions of the two patterns. In the first case, 80% stars and 20% tours can be adequate, in the second one more tours are expected. A weight of zero for one pattern is even possible. The specific proportions can be given from available statistics or can be obtained from available customer data. A method how the ratios can be estimated from traced routes is introduced in detail in chapter 2.3.3.

2.3.2. Default values

As a start for the determination of the variables summarized in MP, a set of default values shall be introduced. Ideally, statistics or customer data are available

which can be evaluated like described in the following sub-chapters. However, often no such data is available and expert knowledge on the industrial sector or good guesses based on experience are required. The settings in this chapter are chosen in order to translate the results back to the classical distance classes and can be used in a general simulation without further input. Due to their interdependencies, the combination of pattern P, *classlimits* and *classcharacteristics* shall be considered in common.
Based on the subdivision into city, local and long-distance traffic, three distance classes are used. The smallest one contains trips with a maximal linear distance of 20km from the base location of the vehicle to single destinations. Inside this radius, all relevant points are of same importance and are thus chosen with the same probability. It is expected that this choice covers larger towns as well as villages belonging together. The next class mainly representing local traffic, gets an upper limit of 50km. In order to clearly separate the classes, only customers settled in a linear distance larger than 20km from the headquarters are accepted here. For the long-distance traffic the two patterns are distinguished. They both take 50km as a lower limit, but the upper one is set differently. For the star pattern, a larger area of operation is accepted since single trips are independent of each other. It does not matter, in which directions consecutive customers are located. Additionally, longer routes are expected to exhibit a smaller deviation between linear and driven distance. A detailed investigation on this topic is given in chapter 5.2.1. In the case of tours a smaller radius of action is assumed to be adequate. It is fixed to 100km since all stopovers have to be visited on one route. The specific locations of two successive destinations can cause large journeys. The extreme situation can easily be understood thinking of two points placed at opposite geographic directions. We then rather expect that the customers are supplied by different vehicles or on different tours.
Further distance bounds of 1,000km and 500km respectively could be included if the considered region is rather broad. Then three classes introduced might not be sufficient to create trips traversing the complete area. An example for such a scenario is a simulation in Brazil. Here, even stars of 1,000km radius cannot cover routes from east to west or north to south. Surely, such trips might be improbable for the vehicle population as a whole, but they can be feasible for single industrial sectors. Since the default values shall be suitable for a large amount of use cases, the partition into three classes is usually sufficient. Another important point referencing the size of the considered region is also its “negative” influence. Like a large area requires larger distances, a smaller one reduces them. Usually, regions have no circular shape such that points lying in acceptable distance to the base location might fall outside the region. Then they are not incorporated in the selection process. Consequently, some distance classes decrease to lower maximal distances naturally or even can be dropped completely for rather small regions.
The *classcharacteristics* consist of the following quantities: *numbers of stopovers* are assumed to be constant and thus described by one value for each combination of distance class and pattern. Here, the default values are rather simple. They

are set to decrease with growing maximal distance. Then a truck conveying goods over a long route performs less trips than a distributor serving businesses in a more bounded region. For the extrapolation to a given target mileage at the end of the evaluation, this shortens the difference in the overall distance traveled in the three classes. The smaller ones add shorter single trips but in exchange a larger number of them. In particular, numbers of 30, 15 and 10 stopovers are used in order to obtain a sufficient overall covered distance. Thinking of postmen for instance, these counts still seem to be by far too small at a first glance. Deliverers are expected to submit more letters a day. However, they do not stop at every house, but also cover small ways by foot. Additionally, it depends on the aim and analysis of simulation results if the number of stops is relevant. In the applications examples given in the following and in chapter 6, in most of the cases one is interested in topographical factors. These stay the same for a journey, no matter how often the vehicle stops in between. Only if the driving behavior has an influence, the results change according to the given number. Furthermore, the simulation might also select one destination multiple times. This usually happens, if there are only few candidates of stopovers in a region. Depending on the pattern, this results in a natural deviation from the default number or in a higher importance of some destinations. Both cases happen in reality and enhance the results instead of falsifying them.
Now, the last information given in *classcharacteristics* is considered. It concerns the direction in which stopovers are chosen. For the default settings, no preferences are set. Thus, the shape of the region prescribes the distribution of customers in a natural way. In general, they can be located in the complete circle or circular ring around the base location without preference. However, if the base location is lying near a border of the region, only few or even no feasible points can be found further in this direction.

Pattern	Star			Tour		
Upper value of *classlimits*	500km	50km	20km	100km	50km	20km
Number of stopovers given in *classcharacteristics*	10	15	30	10	15	30
Proportion in simulation	20%	40%	40%	10%	20%	70%

Table 2.1.: Summary of default parameters, compare to [22]

As a last setting, the proportions of patterns and distance classes are prescribed. Again, these depend on the considered region and industrial sector. For the general simulation, both patterns shall be of the same importance. Their fractions are fixed and shall not be treated as relative frequencies allowing some variation. A summary of the default values is given in table 2.1. For each pattern, a different partition is set. For stars, city and local traffic are of same importance, long-distance traffic is rated less. For tours, the smallest class is clearly preferred

whereas the largest category is reduced. Again, the proportions are prescribed and not used in a frequency distribution. Surely, this can easily be incorporated in a further step. However, for the sake of simplicity, this basic parameter set is kept. It is denoted as $\widehat{MP}$ in the following.

2.3.3. Evaluation of available customer data

A better approach than using standardized default values is of course the employment of customer data measured for the user types of interest. With help of this, the model parameters can be tuned to reflect the vehicle usage in the field optimally. In an ideal situation, starting points and stops are given by their coordinates. After the identification of the vehicle's home location, the pattern of each trip can be determined easily. Therefore the complete trace is split into parts starting and closing at this base, first. Afterwards, every track is identified as one ray of a star if only one stopover is found. Successive rays are combined to one complete star, their number is representing n. Multiple stops approached before returning home indicate tours. Their count again gives n. Furthermore, the distances between origin and every destination are computed. The largest and smallest one allow the assignment to a distance class. If those have a large variation for a star, a split into multiple shapes containing less rays shall be taken into account.
The resulting distribution of pattern, distance classes and number of stopovers is then included in *classlimits* and *classcharacteristics* and the simulation can be started. Taking the proportions as relative frequencies and applying random number generators for the computations of IV, the resulting trips model the customer group very well. This approach has the advantage, that the distances do not have to be measured on the road network but can be computed as linear ones just connecting the points. If the data is anonymized by translation or rotation, the results still are of good quality. We assume that the trips are short enough to avoid problems concerning the earth's curvature when the coordinates are moved. Nevertheless, this method suffers from some problems. There need to be a sufficient amount of data in order to estimate the distributions reliably. Often, recording driving schedules is expensive or time-consuming such that only few data sets are at hand. Additionally, the trips are highly depending on the industrial sector and might not be transferred to other applications. At least, the classification of locations approached is not possible. Since the data usually is anonymized, it cannot be determined what kind of destinations are important.
Depending on the desired final result, a further classification could be of interest. Returning to the subdivision into traffic classes, the customer data also has to be classified as city, local and long-distance traffic. Therefore either each trip, if the data is mixed from different vehicles or user types, or the whole set of data also gets such a label. Usually, this requires some manual work and is influenced by the industrial sector.

Evaluation of traffic surveys

One possible data source for the type of data currently described are polls regarding motor traffic. In case of the traffic survey "Kraftfahrzeugverkehr in Deutschland" (KiD) [26] for instance, public use files containing information on single trips can be ordered. Due to protection of the personal rights of the study participants, no geographical coordinates are computed. However, standardized descriptions of the origins and destinations are given. Additionally, driven distances covered between subsequent stops are recorded. Furthermore, a classification of the industrial sector is included. From this data, the number of stopovers of each trip can be counted. Here the problem lies in the identification of the base location and in determining when it is approached. A classification as site of own or foreign company is inserted, but for businesses having multiple places of location, it might be misleading to treat those points as one. It could also be the case that a vehicle returns to the nearest station between the trips. Additionally, it is not possible to find out if some ways recur. If the same distances and destination classifications are repeated, this does not mean that the same trip is performed. Depending on the usage type, there might also be a significant loss of data since only ten routes are reported in detail. If more are conducted, these are just summarized by their number without providing further information on them.
The advantage of the KiD data set lies in its completeness regarding the trip purpose. For each route for instance it can be identified if it was an official or private trip. In addition, the area of operations is indicated such that vehicles only employed on premises can be distinguished from those used locally, from those driven in the surroundings of the company or in complete Germany. This can be used for a split in city, local and long-distance traffic if required. The mentioned characteristics of the trips unfortunately can only be found in combination in the public use file which is not accessible. The reports available online at [26] directly summarize the characteristics independent of each other. Most of the time only two features are compared in tables. However, they still provide information, like mean values of the number of trips performed on a day, that can enhance the simulation, especially if no further data source is available.
Using the distances reported, either in the data file or in the final report, it should always be kept in mind, that driven distances covered on the road network are given. The simulation of S shown in the next chapter is based on linear distances. Thus, a translation between both concepts has to be performed. Therefore detour factors for a region are used. More details on those and how they can be incorporated in the model are presented in chapter 5.2.1.

Analysis of GPS traces

As a second example of customer input data to adapt MP, GPS traces shall be discussed. These usually are anonymized and can only be taken as reliable infor-

mation if the computation of linear distances is possible without large distortions. Thus, the coordinates can be shifted to start at any place on earth and the whole journey can be rotated around this new starting point, but no scaling or transformation of single points is allowed to have been carried out. If these conditions on the data are guaranteed, the evaluation starts with splitting large traces. Depending on the data source, trips on multiple days might be contained in one set and have to be identified first. If some time stamp is available, this could give good hints where to cut. However, a straightforward split at midnight might be wrong since light-duty commercial vehicles are allowed to drive at night without restrictions. It is better to check for immobilization times, but those can also be misleading due to legal restrictions on driving periods. Where necessary, manual processing might be inevitable. Additionally, single trips have to be identified. Again, automation is not always possible, especially when no information on specific starting points is given. Figure 2.4 sketches an example of a track obtained after the split.

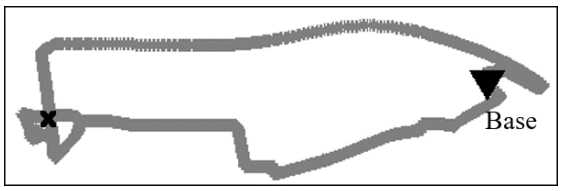


Figure 2.4.: Single tour obtained after splitting, figure taken from [22] was modified.

Next, all routes have to be classified as stars or tours. Usually this has to be done manually by examining each individually. In principle, an assignment using some statistical classification method is possible. However, the training set adjusting the parameters in the corresponding model has to be created first. Hence, a strategy how the trips have to be rated has to be determined in any case. Furthermore, the number of available GPS traces often is rather low. Then, after the required amount of training data is extracted, there are only few trips left that have to be classified. The effort assigning them also manually might be smaller than that setting up the learning algorithm. Returning to the example in figure 2.4, the track is classified as tour. The loops in the west are rated as small detours for stopovers. On the other hand, especially if the base location was set at cross-over point $\mathbf{x}$, also a degenerated star-shape consisting of two smaller and one large rays could be determined. The tour seems to be more probable, but the alternative is not impossible. Obviously, the classification of trips without further information is not unambiguous but is influenced by the knowledge and experience of the person conducting the manual preprocessing.
Whatever approach is used finally, as a result all routes are labeled as stars or tours. If the stopovers are marked in the traces, their linear distance to the starting point and their number can be determined easily. Hence, MP can be set

without further problems. If single destinations cannot be determined, the following workaround can be used. For each continuous track, the maximal distance between all given points and the origin is computed. Thus, at least some information on the upper bounds is given and the distance classes can be split according to them. Of course, some deviations due to the road network and possible detours are possible, but they are expected to be of same magnitude for all traces. Thus, just the class limits are shifted but the distribution is not changed. Unfortunately, there is no way finding out how many stops have been made on one trip. Even if some information on times when the vehicle stood still or when the engine was turned off was given, this would not help determining if it was just because of red lights, level crossings or just traffic jams. It can be used as an indicator, but not as a reliable final result. Then, the default values provided in $\widehat{MP}$ can be used instead as suitable distance limits.

Analysis of GPS data without directions or with reduced information

At last, the special case of GPS data with missing directional information shall be considered. This means, that the data available consists of points that are connected, but it is not clear in which order single points are approached. A minimal example of this type is shown in figure 2.5. It contains ten points where the base location is labeled with A and point C is visited twice. Obviously, the most likely pattern is that of tours. However, it is not clear if path ABC or CBA is included for instance. Thus, the split into two tours is not unique. Figure 2.6 shows three

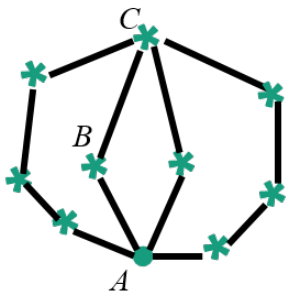


Figure 2.5.: Minimal example for GPS coordinates with undirected connections and marked base location A.

possible ways to assign the single points, the first two options even use identical directions to connect all destinations. If such kind of data is at hand, either with or without the orientation of the arrows, still some important characteristics enhancing the default values can be obtained. Foremost, the distribution of patterns can be estimated like already explained in the sections before. Secondly, the maximal linear distance on a track with respect to the base stays the same independent of the order of the sequence. If $dist(A, C)$ is maximal, it does not

matter if it is computed on route ABC or CBA. Thus, the resulting distance class stays the same. The only thing that cannot be determined without further information is the number of stopovers. Here, two splits with five stops on each tour and one with three and seven are sketched. Consequently, only a mean value or some other estimation can be used. The default values could also be applied if they seem reasonable in the specific situation.

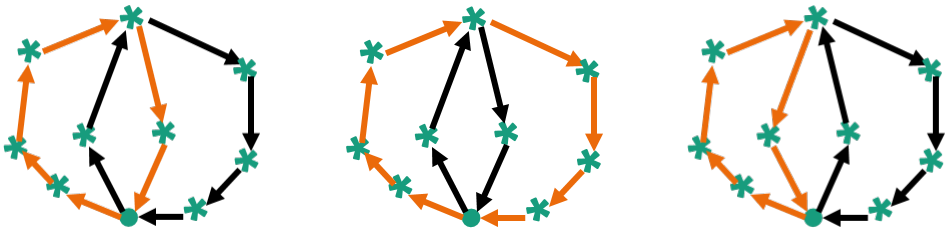


Figure 2.6.: Three possible allocations of destinations to tours, own illustration.

2.3.4. Integration of analyzed customer data

Independent of the specific data source, the evaluation of customer data always results in a distribution of patterns and distance classes. Of course, the number of stopovers is also obtained, but this shall be neglected at the moment. More interesting is the beneficial exploitation of the categorical data. As already noted, it can directly be utilized in the usage simulation, but sometimes results are required for traffic classes or other further partitions. Thus, a transformation between both partitions is needed. In the general paragraph of section 2.3.3, it was already mentioned that single routes or complete data sets can also be enriched with labels for traffic classes. Thus, the routes can be regrouped according to a combination of both divisions. An example for a possible result is given in table 2.2.

Pattern	Star			Tour			Other
Maximal distance	500km	50km	20km	100km	50km	20km	all
Long-distance traffic	43%	1%	0%	52%	1%	2%	1%
Local traffic	3%	39%	3%	6%	47%	2%	0%
City traffic	0%	0%	6%	0%	0%	93%	1%

Table 2.2.: Possible results for the different traffic classes

Here, the composition of traffic classes regarding patterns and distances is summarized. The percentages are computed according to driven distances in the single cells. As an example, the city traffic shall be considered. The corresponding row gives the information that of all analyzed customer trips, 93% of the overall driven distance in city traffic is conducted in tours and only 6% in star pattern. 1% cannot be classified properly. Usually, this results from short trips on factory premises when multiple arriving vehicles have to be organized and parked. Additionally, journeys with missing information, for instance if there was some problem with the GPS signal, are listed here. Those insignificant or problematic routes are excluded from the simulation and their proportion is skipped. Hence, the composition of city traffic is recalculated as 94% against 6% for instance. If the fraction of unclassified trips is high, the feasibility of the data should be reviewed.

2.4. Determination of IV

The set of input variables IV summarizes the settings chosen for a single vehicle. Once the values for B and S have been simulated, the requested route R can be assembled. The specific values in IV have to be determined one after another in the given order. Hence, first U has to be fixed. For the light-duty commercial vehicles considered here, we mainly concentrate on craftsmen and delivery services, i.e. U assumes only two values. "Ideal" examples of them only perform trips in one pattern. Of course, this is not always the case in reality. Thus, depending on the chosen value $U = j$ the probability for a value $P = i$ varies, some conditional probability $Pr(P = i|U = j)$ is prescribed in variable $classcharacteristics$. A random number generator is used to simulate j first, then also i depending on this result. In the default parameter set, only one value i is feasible for each j.
In the next step, the actual distance class is selected. Again, randomness is included via $Pr(D = k|U = i, P = j)$ and a specific value has to be chosen accordingly. Using $\widehat{MP}$, this step reduces to a simple choice depending on the actual $index$. Since the proportions of D are set to be fixed, the complete amount of vehicles can be split directly into groups.
Afterwards, the number of stopovers n is computed. Using traffic polls or GPS traces, information on the maximum m and mean value mv of customers visited on one trip can be determined. Based on these, the probability $\tilde{p}$ of the binomial distribution $B(m, \tilde{p})$ can be estimated with $\tilde{p} = \frac{mv}{m}$. For each vehicle, n can be simulated as number of successful trials from this distribution then.
After those rather simple values have been found, the more complex simulation of geographical coordinates for B and S begins. It starts with the selection of reasonable data. Subsequently, specific points are identified.
The resulting trips shall be routed and evaluated on the road network in the next step. Therefore, they have to be found directly in real map data. This data is

split into two categories, points and areas. Exact descriptions of both types are given in detail in chapter 4.1 where also the methods used to select specific coordinates are explained. Here, only the basic difference and possible values shall be mentioned.

- **Points of interest (POIs)**: POIs are classically used in navigation systems. They specify precise places by their coordinates and give a hint on the use of the marked building. With regard to usage modeling, POIs tag locations at roadside that are worth stopping at. Common examples relevant for light-duty commercial vehicles are fuel stations, supermarkets or pharmacies for instance, as customers of delivery services.
- **Regions of interest (ROIs)**: ROIs are the area-like counterparts to POIs. They identify complete areas experiencing a special land utilization. Examples for those are residential or industrial areas. For certain types of craftsmen, e.g. stopovers correspond to private households, i.e. are located in residential or mixed areas.

The usage simulation for light-duty commercial vehicles is now based on a preselection of POIs and ROIs, differing also between B and S. Here, we still consider a general simulation and thus use the most available data possible. If a specific industrial sector shall be considered, a specialization reducing the relevant data has to precede.
The first location that has to be chosen then is the base location B of the vehicle. Since we are considering commercial vehicles, they are expected to be parked in some build-up area. Their owners' offices are usually placed in industrial, commercial or retail areas. Since family businesses like small crafts enterprises often are run from home, also residential neighborhoods are considered. Inside all of those individual districts, a single place is chosen by the methods described in chapter 4.1.3. In doing so, only the sizes of the areas are used to weight them. Thus, performing multiple simulation runs, more locations found for B are settled in larger districts than in smaller ones. Of course, the proportions might not be reflected exactly due to the use of a random number generator, but the trend should be visible already for few computations. Sometimes, warehouses, distribution centers or even the businesses offices are tagged in maps directly. Companies want to be found easily by potential customers and are registered including their industrial sector. Then they can be used as POIs and B is determined from this data source straightforwardly like it is explained in chapter 4.1.2.
In the last step, the list of stopovers S has to be found. The n geographical coordinates have to be chosen within distance class $D = k$ around B. Again, the methods described in chapters 4.1.2 and 4.1.3 are applied. Prior to that, the database has to be reduced to relevant entries. In the general case, nearly all POI categories like different types of shops, pharmacies, restaurants or post-offices are suitable. Additionally, customers can be located in all types of build-up areas

including commercial and residential ones. Only a specific industrial sector or specialization of the delivery service gives further constraints. The n requested stopovers can then be determined independently. It has always to be kept in mind, that a reduction of accepted data can impede the simulation especially when small distance classes are used. Sometimes, less then n POIs are left, thus at least one is chosen multiple times. This might be correct, but it also might be wrong. The consequences of a restart of the simulation following then are considered in chapter 5 for a special case.

2.5. Setting up R

Having determined B and S, the simulation part is completed and the resulting route can be put together. In the case of a star pattern, this is rather easy and can be done immediately. Since the stopovers are independent of each other, they just have to be alternated with the home location. For tours, further work for reordering the customers is required to find a reasonable driving schedule.
When multiple locations are visited on one trip, the driver usually wants to use some optimal way. The optimality criterion could for instance be the minimal overall distance or the travel time. Some delivery services also apply some strategy that they prefer to turn right which enhances both factors mentioned. Summarizing, the composition of R reads

$$R = \begin{cases} (B,\ S(1),\ B,\ S(2),\ B,\ ...,\ S(n),\ B)^T \in \mathbb{R}^{(2n+1)\times 2} & \text{if } P = \text{"star"}, \\ r = (B,\ r(2),\ ...,\ r(n+1),\ B)^T \in \mathbb{R}^{(n+2)\times 2} & \text{if } P = \text{"tour"} \end{cases} \tag{2.5}$$

where r is the result of a traveling salesman algorithm applied to $p = (B,\ S,\ B)^T$, starting at B. This means that all locations approached on a trip are brought in a clever order such that the linear distances between them are minimized. Afterwards, the tour is shifted to start at B.
The solution of the computations does not have to be optimal. Depending on the specific coordinates, the algorithm might cause disproportional computational effort if the optimum is requested. However, since the distances are measured on the earth's surface without respecting the road network, this "linear" optimum might not be the true one for the real driven route. Additionally, not all vehicles have a tour planning software on board. Thus, the driver's experience and knowledge on the area influence his route. Local drivers usually know shortcuts and time-consuming bottlenecks they try to avoid. This experience cannot be quantified in the simulation. Consequently, there will always be some deviation from the optimal tour in reality and the traveling salesman algorithm can be stopped after having found a reasonable order of the stopovers. Due to the accepted discrepancies, the distances could also be estimated using plane geometry and the

Pythagorean theorem instead of calculating them on the earth surface. We are aware that for larger distances they can deviate drastically but the saving of time is rated higher.
In a prospective version performing these computations in VMC® , real distances could be used in the algorithm enhancing the results. However, the computation time might grow drastically if no precomputed distance matrix is available. Then ways between all coordinates, base location and stopovers, have to be found. Indeed, then time- and traffic dependent routes can be optimized.

2.6. Processing and evaluating created routes with VMC®

The outcome of the simulation consists of an amount of kml-files (kml= Keyhole Markup Language, a data type to store geographical information comfortably), one for each route R, including coordinates of places visited on one journey each. These are now further processed with VMC®, a software product developed by Fraunhofer ITWM, see [53]. The methods described in this section are implemented and can be applied directly. The only work left to do is the choice of options and parameters provided there.
First of all, the obtained routes consist of geographical coordinates located near the roadside and have to be transferred to tracks on the road network. Therefor the routing algorithm of VMC® is used. First, it projects each point to the nearest suitable road. "Suitable" in this context means, that sometimes streets a bit farther away are preferred, for instance if motorways are passing. This concept is explained in detail in chapter 5.2.1. Afterwards, the search for a connection of successive coordinates is started. Here, the customer specific travel behavior can be incorporated by adapting the optimization parameters. Commonly, the choice between shortest and fastest routes is included in these settings. Depending on the area of operations the vehicle driver for instance might prefer the first option saving driven distance and thus reducing fuel consumption or the second one if goods have to be delivered in a certain time slot.
Afterwards, the traversed roads are analyzed regarding different aspects. Factors to compute can for instance be related only to topographical data like hilliness, curviness or slopes measured on single road segments differentiated by their type. This can for example be based on the highway types used by OSM like motorway, major, primary and so on but can also be adapted to user defined designations. The separation between urban and rural sections is also of huge importance. As a result, for each trip a table summarizing the computed values for each road segment is obtained. Figure 2.7 shows an extract of the export of such a table. It includes the location of each segment, its road type, length, etc. Its specific appearance depends on the factors chosen and the settings of VMC®.

altitudes	coordinates	countrycode	length	roadtype	settlementarea
105,104.7870000000	LINESTRING(8.6145198 49.4667	DE	0.880514000	Stadt B	urban
102.492,102.916,103	LINESTRING(8.6067896 49.4710	DE	0.563686000	Stadt A	urban
101.996,101.053,102	LINESTRING(8.5996671 49.4730	DE	0.162049	Land A	rural
102.116,101.938,100	LINESTRING(8.5976317 49.4736	DE	0.363815	Stadt B	mixed
101.645,100.057,100	LINESTRING(8.5960336 49.4758	DE	0.152570000	Stadt A	urban
99.535899999999998	LINESTRING(8.5950178 49.4746	DE	0.863655995	Land A	rural
102.708,100.858,101	LINESTRING(8.5851047 49.4788	DE	0.153259000	Stadt A	urban
100.175,99.36079995	LINESTRING(8.5836334 49.4795	DE	108.833	Land A	rural
99,100.158,99.59470	LINESTRING(8.5718712 49.4743	DE	0.674849000	Stadt A	urban
98.884699999999995	LINESTRING(8.5655834 49.4701	DE	0.283262000	Land A	rural
100.31699999999995	LINESTRING(8.5672379 49.4682	DE	0.0267166	Stadt A	urban
99.933199999999995	LINESTRING(8.5672935 49.4680	DE	0.215851995	Land A	rural
97.509900000000002	LINESTRING(8.5649483 49.4668	DE	0.231652	Stadt A	urban

Figure 2.7.: Extract of the segment table exported from VMC®.

Here, we consider the same outputs as requested in [22] and [60]. The factors of interest are hilliness and curviness, both split in two categories, which are combined to the three topographical classes low (flat and straight), high (hilly and curvy) and moderate (else). More details on this choice are given in the second reference. Summarizing all segment tables for all simulated vehicles gives a table like 2.3. Again, the proportions are computed based on the driven distances conducted in each combination of pattern, distance and road type.

	Road type					Topography		
Pattern	Motorway	Rural		Urban		High	Moderate	Low
		A	B+C	A	B+C			
Star 500	79%	10%	1%	8%	2%	34%	46%	20%
Star 50	44%	11%	10%	25%	10%	22%	45%	33%
Star 20	20%	10%	15%	20%	35%	24%	40%	36%
Tour 100	50%	17%	7%	18%	8%	31%	41%	28%
Tour 50	25%	13%	14%	33%	15%	33%	33%	34%
Tour 20	18%	12%	22%	17%	31%	26%	39%	35%

Table 2.3.: Exemplary results for the created routes.

2.7. Combining results

Finally, the obtained results can be combined and used as an input for a usage simulation estimating damage values.
Though first of all, tables 2.3 and 2.2 are merged. The former shows the distribution of requested topographical factors based on patterns and distance classes, the latter links this partitioning of routes to the requested one into traffic classes. A proper matrix multiplication of both tables easily gives the desired result shown in table 2.4. The traffic classes are analyzed regarding the chosen factors. Often the final outcome shown in the last table is not satisfying. A further estimation of load distribution is desired. This can be simulated based on the already provided data in combination with results recorded at measurement campaigns.

	Road type					Topography		
Traffic class	Motor-way	Rural		Urban		High	Moderate	Low
		A	B+C	A	B+C			
Long-distance	88%	4%	4%	3%	1%	31%	46%	23%
Local	45%	11%	12%	14%	18%	30%	48%	22%
City	11%	13%	19%	17%	40%	32%	47%	21%

Table 2.4.: Combination of results for customer simulation, taken from [22].

2.7.1. Planning and evaluating measurement campaigns

Planning a measurement campaign is a challenging task since a balance of the quality and completeness of results, costs and further conditions has to be found. A detailed description of the process of planning, performing and analyzing measurement campaigns is given in [60]. Here, only the basic concepts shall be explained. The goal of a measurement campaign in this context is the determination of road conditions in a specific region and their influence on the vehicle. Therefore, data like the forces on different components, acceleration or other interesting information is collected and stored in combination with positions determined by a GPS receiver.[22] In order to achieve a reliable statistical distribution of all relevant elements the driven route has to be chosen smartly. All combinations of factors of interest, like topographical ones or the road surface for instance, have to be measured with sufficient accuracy. On these grounds the region as a whole has to be examined first, then a matching route has to be found well reflecting the obtained distribution of factors. On the other hand, the entire trip is not allowed to take too long. The fixed costs of bringing the measurement vehicle to the considered region, as well as the time dependent ones like personnel costs of the drivers and

crews, must not exceed a predetermined bound. Usually these limits are formulated as time conditions on the campaign. Thus, especially for broad areas, the proportion of small roads is reduced in favor of motorways where large distances can be overcome faster. This enables measurements in more parts of a region and enhances the representativeness for the conditions in the country as a whole. In doing this, not only cities with good transport connections and expected better road conditions but also more isolated villages with possibly older paving can be reached and included in the sample.
After the campaign has been conducted, the measured data has to be analyzed and checked for errors and implausible values. Afterwards, the remaining GPS traces are again evaluated with VMC®. First, the route is projected to the map and segmented. Afterwards, the same factors like those used for the simulated routes are computed. In addition, each segment is also equipped with the values measured during the campaign. Since the campaign often is conducted only for one load setting, further calibration measurements are performed. These allow an estimation of the forces for different loads. Details on this procedure are given in [60].

2.7.2. Simulating customer specific load distributions

The damage values for single users are then simulated with the software U·Sim, also developed at Fraunhofer ITWM. It takes the division of factors from usage modeling and measurement campaign as input values. According to the first distribution, segments from the campaign combined with their measurements are drawn until a desired target mileage is reached. Then the expected damage value for this trip is computed. For each traffic type a large number of customers is simulated this way and damage values per kilometer are computed. Based on these, the three groups can be compared to each other. For a sufficient amount of data even reliable quantiles for the damage values for single measuring channels can be computed. The complete process is explained extensively in [54] and [60].

3. Usage models for passenger cars

3.1. Introduction

A further usage model that shall be considered is that of passenger cars. In contrast to the model introduced before, here private and commercial use has to be differentiated. Both types can also be characterized by $R = R(GI, MP, IV)$, with MP adapted to the specific case and slightly different interpretation of the components of IV.
The main user group of private passenger cars are commuters. Their usage model has already been introduced in [21], summarizing the main facts. Here, we want to have a detailed look at all facets of frequently performed trips. Taxis form an important part of commercial users of passenger cars. They show much more variation in their routes since they have to react on their passengers. They often perform many trips and cover large distances during one shift.

3.2. Usage model for private passenger cars

The most mileage of typical commuters is covered during workweek. Thus, we restrict ourselves to the simulation of five typical days. We skip journeys going on vacation and routes on weekends in order to keep the model simple. These two exhibit a great variation requiring various input to compute representative routes. In addition to the regular trips between home and work place, different kinds of private trips are included. These contain more or less regular shopping trips to the supermarket or leisure time activities.
In principle, it is not important for the simulation when trips are performed, thus those scheduled after work can for instance be adopted to happen on Saturday. If only the topographical characteristics of the roads traveled are relevant, the order of tours can be completely neglected. Only if influences of the motorist, like his manner of driving, are of interest, the starting time can be important.

3.2.1. Characterization of model

The usage model for commuters can be described by the same structure like that used for light-duty commercial vehicles in equation (2.1), split in three parameter sets. GI stays the same for all usage types. The region where routes shall be generated and a unique index identifying single vehicles are always required. The model parameters MP are more extensive. Since different types of trips have to be simulated, multiple distance distributions have to be provided in *classlimits* and *classcharacteristics* (compare equation (2.3)). Furthermore, the transition probabilities between tours of different classification are needed. A new parameter *partitions* is inserted to store parameters of additionally required distributions independent from those for distances. IV contains the same components as in the previous model. They have the following interpretation:

- $\boldsymbol{U}$: The only private users of passenger cars considered are commuters.
- $\boldsymbol{P}$: The driving pattern for commuters is distinguished in three groups differing in additional regular trips. One performs none, one has a stopover attending children and the last goes out for a trip during break. A description of all of them is given in section 3.2.2. Due to the influence of the cultural background, the expected frequencies for the patterns are included in *partitions*.
- $\boldsymbol{D}$: The single distance class sufficient for light-duty commercial vehicles is expanded to a set of distance classes for different types of trips. At least it contains a choice for the distance between home and work location and one for an activity like the weekly shopping. The probability for a single entry $D_i = k$ depends on the chosen pattern and trip purpose, $Pr\,(D_i = k\,|P = j, triptype = m\,)$. m takes feasible values provided in *classcharacteristics*.
- $\boldsymbol{n}$: Number of leisure time activities. This variable determines the number of trips performed additionally to those prescribed by P. Up to now, this number is taken as a constant value set in *classcharacteristics*. Thus, each commuter simulated conducts the same number of additional journeys per week. In a further step, more variation shall be added. If the distribution of the count of such trips is available, n can be enlarged to a vector distinguishing the types of destinations like shops selling convenience goods or sports facilities.
- $\boldsymbol{B}$: Home location of vehicle. Since private passenger cars are considered, the base locations of those vehicles usually lie in residential neighborhoods inside *region*. The rare case that a garage or other parking lot used regularly is farther away is skipped in the model. It can be incorporated by inserting

additional trips in the evening, but it is assumed that the enhancement of the result is not worth the effort.

- $\boldsymbol{S}$: List of destinations. S gives the coordinates of all places visited except B. At least it includes the work location and one leisure time activity. Depending on P, further entries have to be provided.

The route R is then again assembled from the coordinates given in the last two entries. Here, not only one driving schedule is created but five, one for each day in workweek. Some of them are the same since P is constant but the additional destinations are distributed over the week. The complete process is shown in section 3.2.5.

3.2.2. Simplification of model

Daily routes between home and work form the most important trip during a work week, but also additional private tours are not negligible. They can be divided in different types contingent of order or frequency. The partitioning which plays the larger role is when additional destinations are approached. Here we can distinguish six cases that are sketched in figure 3.1. First of all, stopovers can be included in the way to work. Typical examples for this situation are parents dropping off their children at kindergarten or school. The next possibility for tours is during break when people go out to get food or take care of quick errands like going to pharmacies. Private matters that take longer, like going to doctor or doing the weekly shopping, are shifted to the way home or are executed on new trips starting from home. The latter can be conducted either after work but sometimes also before work. They are assumed to end at home again.

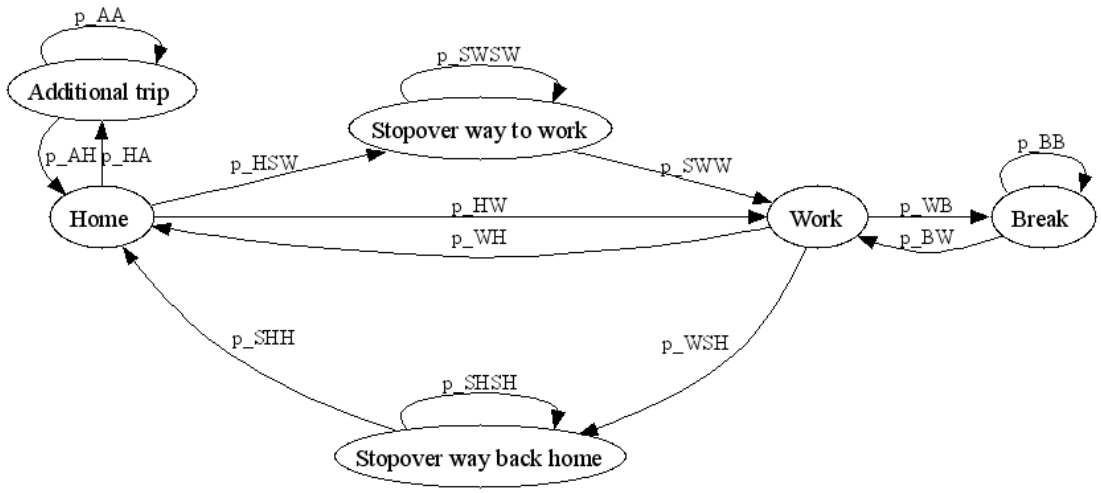


Figure 3.1.: Time schedule for additional target points

Since the simulation at this point does not respect starting times, these two cases can be handled as one in a simplified model. However, more important is the concentration on three base patterns. The first and most relevant is that of a commuter who mainly travels between home and work. In contrast to him, there exist people having a stop on the way to work and partly also on the way back. They have a fixed routine bringing their children to school or kindergarten and collecting them again. The last and smallest group are the persons exploiting their breaks. All other types of stopovers and additional trips are included in the simulation for every driver.

3.2.3. Setting up MP

classlimits and *classcharacteristics* have to provide distributions of driven distances for each trip type. Furthermore, the target points can be categorized by their frequency of occurrence. This depends on the chosen driving pattern. Some of them, like kindergarten and schools, are visited everyday, partly before or after work or on both ways. But those are only relevant as additional stops for one commuter pattern. Leisure time activities are performed multiple times a week, depending on the number and type of hobbies. There are shops of daily or weekly needs as well as stores that are visited once a month or even less frequent like home-centers. All of those stopovers are of interest for everybody.
Unfortunately, the transition probabilities p_{ij} indicated in figure 3.1 are not constant for individual drivers. They always depend on the complete route the considered person already traveled. If one or more supermarkets have already been visited on the way home, it is less probable that an additional way to a further store is conducted. If the person went straight home, it is more likely. This probability even grows, if on previous days just the basic trips were performed. That makes the simulation more complicated and the theory of Markov chains cannot be applied. Therefore, a constant number of additional trips per week is prescribed in the simplified model at the moment. Thus, n always stays the same for each commuter but a variation between the days is included, like shown in section 3.2.4. Of course, even for the small number of two or three extra trips per week, there does not always have to be only one stopover on a day. The selection is done independently using a uniform distribution. Multiple destinations are just approached one after another. On the remaining days then only the basic trips are performed.

As already mentioned, different distance distributions have to be included. Those vary in the underlying source data and their extensiveness and are incorporated in differing ways. The routes to work and to leisure time activities can be based easily on available statistics. The stopovers for attendance and during break are treated differently:

Distances between home, work and general leisure time activities

The way between home and work is the major part of the weekly driving schedule. Thus, a reasonable choice is essential. The main constraint the workplace has to fulfill is its "suitable" distance to the place of residence. Of course, it should not be too large, since people only accept a certain distance and travel time. If they are unacceptable, people move, live in a nearer place during week, or search for a new job, depending of their social and familial situation. On the other hand, the daily traveled distance should not be too short, otherwise people prefer going by bike or just walk. The limits in both cases depend on individual preferences and mainly also on the infrastructure of the region they live in. Persons originating from rural areas rather tend to go by car than inhabitants of large cities accustomed to a well developed urban transport system.
The process of selecting a workplace is contingent of the data situation. The results of mobility surveys like "Mobilität in Deutschland (MiD) 2008" [38] or its French counterpart "Enquête nationale transports et déplacements (ENTD) 2008" [46] are used as a basis. In both, people are asked to report all trips on one randomly selected day. These mobility diaries include not only driven distances but also the vehicle used, the trip purpose and further personal or geographical characteristics. The designated sources provide evaluations in tabular form. They contain distance bounds and corresponding frequencies for the selected regions. However, it is of great interest to determine influencing factors for these distances. Then the mentioned differences between rural and urban areas can be quantified. Here, more detailed driving schedules including information on the region of the home place are required. The analysis of the German data and its preparation for usage simulation is described in chapter 4.3.1.
The obtained distance distributions, independent if they result from tables or more verbatim reports, still exhibit of one problem: People naturally register covered distances measured on the road network. In the algorithms used to simulated the entries of S we can only work with linear distances. A straightforward solution to solve this problem is the application of so called "circuity factors" introduced in [18]. They quantify the necessary detours between both distance calculations and can be used directly to translate them. However, the values proposed are estimated only from a small number of road segments and summarized to one value per country, though they should be dependent on topographical factors. In chapter 5.2.1 an algorithm enhancing the provided detour factors is shown. They can then be applied to perform the calculations in section 5.2.2 computing the distribution of driven distances. As a result, a list of intervals with respective frequencies for the required trip type in the considered region is obtained. These are stored in MP then.
This procedure can also be applied directly for the distance distributions for general leisure time activities. They can be included in a similar way like just described. Depending on the richness of detail of the simulation, these distributions

can be further separated regarding the classification of destinations. For each, additional entries in *classlimits* and *classcharacteristics* including frequencies are required. In the most simplified model, only two distributions, one for the way to work and one for distances to stops measured from home, are provided.
It is important to keep in mind that in the summarized results the connection between trips is lost. Hence, a stop when returning home can split the way in one part from work to a supermarket for instance and the other one from the supermarket to home, but both distances cannot be linked anymore in the table. Thus, if for either section a distance class has been chosen, these might not fit together. If no feasible destination fulfilling both displacements can be found, the simulation restarts. In order to reduce this waste of effort, a single suitable distance only is accepted. Since the home location is the central point of the simulation, the distribution of trip length between shopping and home is rated higher. Surely this can lead to detours if the stopover is not directly on the way, but people usually accept larger distances for destinations worth it.

Distances for trips attending children and during breaks

The ways that have some special characteristics are those attending children and those performed during break. Both of them are not tagged in a directly usable way in the survey data. The former can be obtained by comparing the traffic diaries of adults and children living in one household. On one hand, it has to be found out which persons were traveling together, on the other hand, the purpose of the trip has to be selected correctly. If children are accompanying their parents, the true value is given in the adult's diary, if they are dropped off at the kindergarten, parents have "attendance" as main scope. A method to combine records belonging together in order to obtain feasible distance distributions is given in chapter 4.3.1. This evaluation is only possible if this special type of data is available. Usual statistics do not contain the required information. Then a different approach is more suitable: The distribution of kindergartens and schools strongly depends on administrative structures and local conditions. Usually, the nearest or best accessible institutions are preferred. Thus, it is assumed that a selection in the same commune or in a certain radius provides adequate results with less work.
The same holds for trips during breaks. They are not classified with a special trip purpose but are assigned to the standard types like shopping or detailed leisure time activities including going out for lunch. Hence, the order of all destinations has to be reviewed and those between two trips ending at work are marked as approached during break. Details of this procedure are also given in chapter 4.3.1.
Due to the restriction on a single workplace omitting people with multiple jobs in the simplified model, the resulting distance distribution might be wrong. An employee traveling between two working places rather covers a larger distance then somebody who has to return to his origin within 30 minutes for instance. Usually,

it cannot be determined from traffic surveys if different places are targeted since the destinations are just classified as "work" without further information. Again, the evaluation is even not possible with summarizing statistics. An upper bound for the distance is expected to be sufficient as well. People only have limited time available and will not drive that far from work.
In this simplified model we use 30km in case of attendance trips and 10km for break activities as default values for maximal covered distances. Especially the first bound might be doubted as being too large, but this is a consequence of two facts. First of all, in the MiD data there are trips larger than 25km reported that should not be neglected. Additionally, simulations conducted with a smaller maximum showed a lot of restarts due to non available candidates for target points. Enlarging the search radius made the procedure more stable. Since the longer trips are not assumed to be too bad and parents often accept big detours for their children, the larger limit is accepted.

Some further settings included in *partitions*

In addition to the distance distributions, some further frequencies have to be included in MP. First of all, the proportions of the three commuter types have to be determined. Here, traffic surveys could be used again. Available driving chains have to be checked for attendance trips or activities between two ways to work, see chapter 4.3.1. If any of these special routes is included, the commuter pattern is clear. All remaining persons are summarized as the main group driving to work and back. As default, frequencies of 20%, 10% and 70% can be taken, see [21]. The parents additionally are split into two groups performing one or two extra trips. The proportion of attendance trips strongly depends on the region considered, since the cultural background influences the attitude of parents. In some countries for instance children usually go to school by bus whereas in other states parents prefer bringing them to school themselves.[21]
For the general additional trips also some probabilities have to be prescribed. People can again be split into three groups performing trips on their way to work, on the way home or additional trips. Of course this kind of information can also be extracted from traffic surveys, if available as detailed reports. However, often such data is not at hand or the results are not rated as reliable since only single days are reported, no complete weeks. As already mentioned in the introduction of this chapter, the p_{ij} in figure 3.1 are not constant. Thus, the mean value of number of trips performed on one day shall be used to find a good approximation. In the simplified model, no trips besides that classified as attendance are simulated on the journey to work. The leisure time activities after work are divided into 52.5% performed on the way home and 47.5% as new trips.[21]

3.2.4. Determination of IV

Since for commuters the user type U only attains a single value, the simulation of IV directly starts with the choice of a pattern. Due to the reduction to the simplified model, the prescribed frequencies for the three commuter types can be easily inserted in some random number generator selecting one of them. Concerning n, no work has to be done since this value is taken as constant at the moment. If in a next step more variation shall be added, the mean and maximal count of leisure time activities during a week can be used to determine a binomial distribution and individual numbers can be simulated without much effort.
The determination of D, B and S is more demanding. The number of entries needed in D and S depends on P. Additionally, all values have to fit together. If one selected distance class cannot be fulfilled for the chosen home location, the complete simulation has to be restarted. An example for that situation is a commuter living in some rural area. If a short way to work is selected, there might not be any feasible candidate for the location inside this distance. Thus, that person and the corresponding driving schedule are not realistic.
The order of simulating D and B depends on the distribution classes provided in D. Their frequencies usually are assumed to be constant for the complete *region* such that the choice of an initial interval is independent of B. However, otherwise or if the detour factors required to transform the interval bounds vary, the home location has to be chosen first. The specific procedure depends on the data given in MP. We concentrate on the case that the detour factor is constant with respect to the influence of *region* and stay with the assumption that *classlimits* already contains linear distance bounds. This does not mean that the detour factor cannot vary for different distance classes, see chapter 5.2.1.

Simulating D

The distances classes stored in D vary with the chosen pattern P. In any case, bounds for the way to work are included. These are determined from the corresponding distribution given in MP. Limits for the n leisure time activities are chosen independently from each other for each commuter. Again, a frequency distribution of reasonable intervals is provided in MP. A single interval is chosen randomly applying the cumulative-size method given in [44]. Further entries are only needed for the two special patterns. Since the simplified model uses prescribed upper limits, those only have to be repeated if required.

Choosing a home location B

The choice of the precise place of residence of the commuter is the crucial point since it determines the starting point of the complete simulation. All trips depend

on its location. This origin obviously has to be settled in some residential area of the considered region or country. Hence, the first challenge is to scatter the home places correctly. Thereto the true distribution has to be known or it has to be estimated. Depending on available data, there are multiple possibilities to do this. They are sketched in figure 3.2.

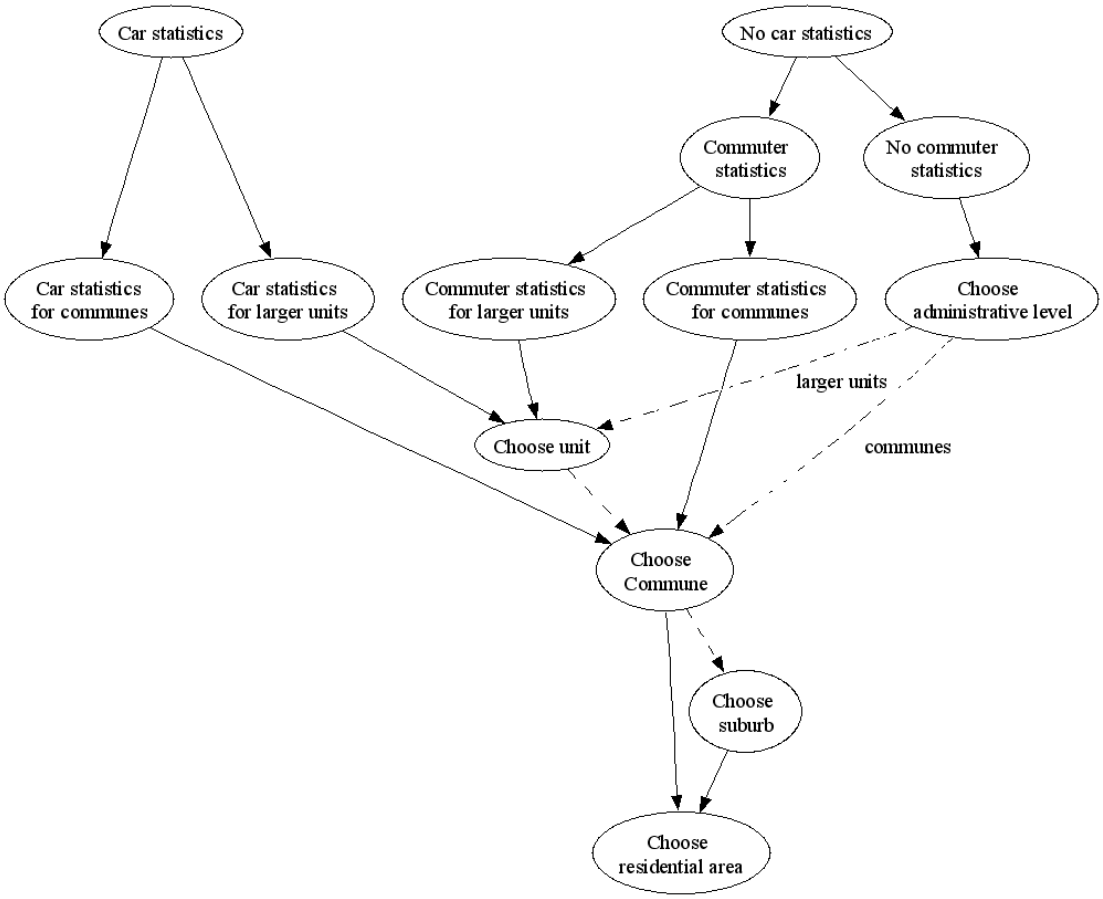


Figure 3.2.: Different possibilities for choosing a home location. "Larger units" are short for rural districts and urban communes or federal states. Dashed lines represent the selection based on population figures.

In an ideal situation, the distribution of private cars is known for the complete considered region. This means the number of vehicles is available for every commune. Depending on the division into further units and the presence of population numbers like those of suburbs, a smaller district can be chosen first or the residential areas are used as basis directly. In the next case, the vehicle distribution is given on administrative units like counties or states. Again, population figures allow the selection of a city or village before concentrating on neighborhoods. Extensive details on the method are given in chapters 4.2 and 5.1. If no vehicle statistics are at hand, two more deficient methods can be applied. At least for Germany, the

population figures have an area-wide good quality up to the rural district level. They can be used to select communes, assuming that vehicles are distributed uniformly to people. This presumption is not completely correct since large cities with better infrastructure concerning public transport are rated too high whereas rural districts, where few people live but proportionately possess more cars, are underrepresented. However, as shown in chapter 4.2.2, the resulting distributions are acceptable.
The special case of available commuter statistics is considered in section 3.2.6. Applying them slightly changes the simulation procedure and shall not be considered here. As a result, also a list of residential areas as candidates for areas containing B are computed.
Independent of the actual determination of these candidates, the choice of home location inside the preselected residential area is performed with the standard method for choices in ROIs described in chapter 4.1.3

Selecting a work location

After the selection of the home, the job location has to be found. We concentrate on commuters with a fixed workplace which is constant over the complete simulated week. We postulate that no business trips, including customer visits, have to be performed with private cars. We assume that such routes always are covered in company cars that shall not be considered here. The working place can be located at various positions. First of all, most kinds of POIs are adequate including shops, fuel stations, restaurants, etc. Additionally, it can be situated in retail, industrial or commercial areas, obviously. Further more, like it was explained before, family businesses often are settled in dwelling houses. Certainly, there are also persons hired not living there. Besides, caretakers or household helpers also work in residential areas. Thus, these also have to be included in the simulation. What is still to consider in detail are large facilities like universities or hospitals. They are often only marked as POIs but are not included in one of the mentioned types of areas. Compared to supermarkets, usually more people are employed there, but this is not reflected in the simulation directly. Hence, we introduce a new class of areas called "work" summarizing such large buildings that are underrated otherwise. Since we do not have any information if POIs or ROIs are more important, we simulate the same share from both by default.
The selection of a specific place is performed by applying the methods given in sections 4.1.2 and 4.1.3. The corresponding entry of D is used to determine the circular rings in which the work places is searched. The result is stored as first part of S.

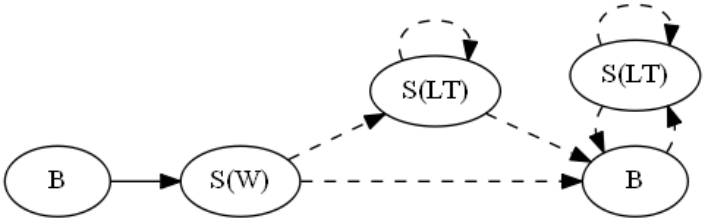
Determining additional stops

The selection of additional stops again splits between leisure time activities after work and those for the two specific patterns. In any case, the destinations are chosen from POIs of multiple types of classification. For trips with purpose attendance, clearly kindergartens, day-care centers and all types of schools are suitable. During breaks small tours going to cash points, pharmacies or post boxes for example are considered. Of course, cafés, restaurants, bars etc. are also included. Leisure time activities can take place in even more locations, though the list is enlarged with all sorts of shops, sports facilities, doctor's offices and so on.
Again, the methods presented in section 4.1.2 are applied. For the additional trips due to the commuter pattern the method selecting a POI in a given radius is used. Certainly, for the break activities the circle is drawn around the work place. All other distances are measured from B. For the general trips performed by everybody the distances classes stored in D give upper and lower bounds. Even if multiple additional destinations are approached on the same day, they are chosen independently from each other. The obtained stopovers are added to S, also recording the trip type.

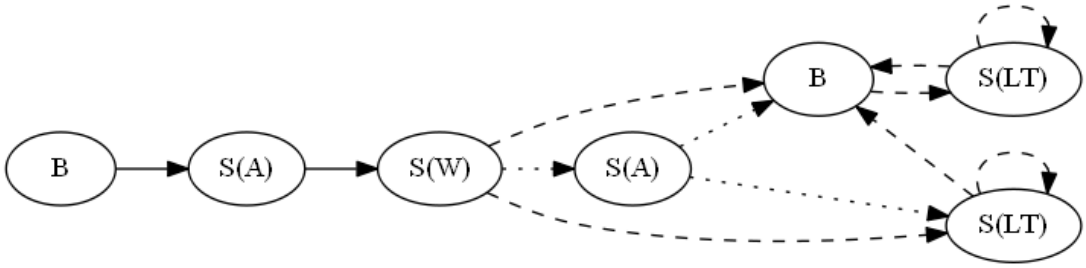
3.2.5. Assembling R

After all relevant coordinates stored in B and S have been found, R can be put together. In contrast to the simulation of light-duty commercial vehicles, here five kml-files, each listing the route points of a single day, are created. They all start with the characteristic part of the commuter pattern. In the easiest case this means home and work are concatenated. For the attendance pattern, the secondary destination is inserted between these points. In case of an activity during break, this target point is appended and the work location is repeated.
The leisure time activities now have to be distributed over the week. Therefor n uniformly distributed integers between one and five are generated. These indicate the days when the additional trips are performed. Then, n decisions based on the frequencies given in *partitions* are made, determining if the stopovers are approached on the way home or if new trips from home are started. The kml-files selected in the first step are then extended according to the decision in the last step. Figure 3.3 summarizes all possibilities. The entries of S got an indicator on the specific trip purpose. S(LT) represents **l**eisure **t**imes activities, S(A) and S(B) respectively characterize the routes attending children and during breaks. Solid lines mark ways that are conducted on every day by each passenger car assigned to the considered commuter pattern. Dashed lines are used if routes are possible but not guaranteed. Looking at the first graph for instance, the work place is approached every day. The leisure time activities are first of all not performed every day. Secondly, they might be included on the way home or form new trips. The sketch of commuters attending children additionally contains

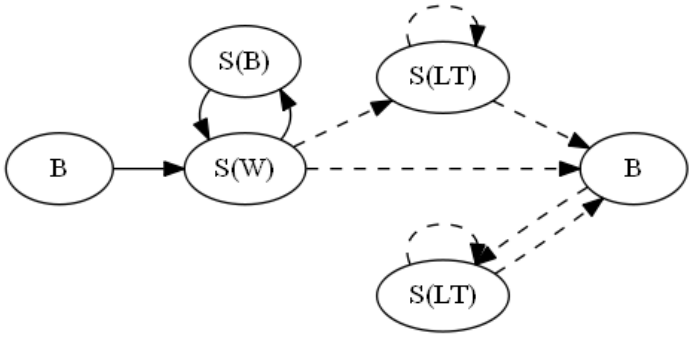
dotted lines. These designate routes picking them up again. Since this pattern is split, a person drives those ways every day in workweek or never. Trip chains belonging together, like the two individual ways from home to leisure time activity and the way back, are styled in the same manner even if only the first one includes the option. The following need to be conducted if the first one was performed. In the given example this just means that people have to return home in case they went out again.



(a) Commuters only going to work and back home



(b) Commuters with additional attendance trips



(c) Commuters performing trips during their breaks

Figure 3.3.: Route components split by commuter patterns

In the outcome of simulating additional destinations, multiple targets can lie on one trip. One possibility would be to order them such that the overall distance of that route is minimized. An algorithm solving the traveling-salesmen problem could be adapted to sort the locations. However, the quality of the result does not compensate the computational effort. On the contrary, it is often even not correct. People choose their route intuitively, preferring ways with short time instead of distance. Furthermore, they possibly do not find what they are search-

ing for in the first store and then decide to go to another. The resulting detour was not planned before but was required and accepted. Thus, the unsorted stopovers are also correct and the additional effort reordering them is not necessary.

3.2.6. Inclusion of commuter flow matrices

One type of statistics for distances between home and work location that has not been considered yet is that of commuter flows. Those summarize the combination of both places in tables, i.e. the rth row and the cth column contain the number of people commuting from area r to area c. Thus, no distance classes are directly available. This type of data can be included in the commuter model by omitting the according distributions in MP and changing the way B and the work place in S are selected.
Commuter flow matrices usually have the administrative units for the places of residence given as rows and those for the work locations summarized in columns. Adding up all entries gives the total t of all registered commuters. Summing up single rows or columns gives the counts r and c split by units. Hence, different types of simulations can be performed.

- First, a unit h containing the home location can be selected based on their frequencies given in r. In order to choose a work location, the counts in row h are taken as partitioning and a unit w is selected randomly considering the frequency distribution.
- The same can be done the other way around. A unit for the workplace is selected depending on c first, then one containing the home is simulated.
- The last possibility is to consider the complete subdivision of t and choose a cell representing both units in combination.

In all cases the selected units have to be matched with their residential areas for the home location and suitable POIs and ROIs for the work places. Then again the methods described in section 4.1.2 and 4.1.3 have to be applied to determine concrete coordinates.
At a first glance, commuter flow matrices seem to be more adequate for the simulation than the vehicle distribution used before, but we assume that they are less reliable. Vehicles are usually registered on the place of residence of their owners and are combined with their types, like private or commercial. Commuter statistics have the disadvantage that workers often are not recorded at their real place of work but at the head office of their employer. Additionally, commuter statistics contain employees using all means of transport, not only passenger cars. If these statistics are used anyway, the available detail of administrative units is also essential. Usually only larger regions are summarized. A reduction to the

lowest level, for instance based on population figures, should be performed before the residential areas are employed.

3.3. Usage model for taxis as commercial passenger cars

The commercial usage of passenger cars is very variated. Some cars are necessary for daily work, like in case of sales representatives or distribution services, others are just used occasionally for instance for irregular business trips. Most of these routes can be modeled like those for light-duty commercial vehicles with appropriately adapted parameters. However, there are customer groups that cannot be included that way. One of them are taxis. They are on the move every day and perform many different kinds of trips.
In the following chapter, a usage model for these special commercial passenger cars is introduced. Seven days are simulated for one vehicle, distinguishing between differing characteristics of workweek and weekend. In this, hailed shared taxi having a rather controlled schedule are not modeled explicitly. They can be included by adding bus stops and a suitable frequency of trips between them.

3.3.1. Characterization of model

As already stated at the beginning of this chapter, the usage model for taxis can also be summarized by equation (2.1). Again, the single components have a special meaning in this case. In contrast to the two usage models already introduced, that for taxis is the most complex and contains more variation between the created single routes. MP includes a multitude of different distance distributions not only distinguished by trip type but also by the distinction between workweek and weekend. The input variables in IV consist of the following components.

- $\boldsymbol{U}$: Commercial users of passenger cars are split in multiple groups like salesmen, distribution services and especially taxis. Only the latter are modeled and simulated according to the methods shown next. For the first examples the algorithms introduced in chapter 2 can be applied.
- $\boldsymbol{P}$: The driving pattern for taxis is only split in two cases. In the first one, some kind of self-employed taxi driver is assumed who can only exercise one shift per day. On the other hand, a vehicle owned by a taxicab company is assumed. Then two separate shifts a day are simulated where the drivers meet at the head office to hand over the car. Apart from that, the second pattern just doubles the first one. The restriction to these two cases is

sufficient to explain the overall concept. Surely also more shifts can be simulated easily.

- $\boldsymbol{D}$: Like it was already introduced for private passenger cars, also the route of taxis consists of different types of trips. Their explicit number cannot be predetermined since it depends on the specific composition. Some types contain multiple tracks of varying length. At least n distance classes have to be simulated, the extreme case requires $2n$. The probability for a single entry depends on the chosen pattern, the trip purpose and the current part of that trip, $Pr\left(D_i = k \,|P = j, triptype = m, trippart = l\right)$ where m takes feasible values provided in $classcharacteristics$ and l is suitable for m. An easy example describing the situation for two parts is the common case that people call a taxi to drive to the airport. Then first the distance between current taxi position and home location of customer has to be selected, afterwards the distance between home and airport has to be chosen. Those are independent of each other but not all values are accepted here.

- $\boldsymbol{n}$: The count of performed trips $n \in \mathbb{R}^{7\times j}$ conducted on each single day, where $j \in \{1, 2\}$ indicates the number of shifts simulated. The entries in the first five rows of n are independent and identically $B\left(n_{max}, \tilde{p}\right)$ distributed. n_{max} indicates the maximal number of trips on a day during workweek provided by some statistic and $\tilde{p}$ is determined such that the also reported mean number of trips is achieved. The same structure is applied for the last two rows, the distribution parameters are computed for the weekend accordingly.

- $\boldsymbol{B}$: The base location of the car is influenced by P. A self-employed driver is assumed to park his taxi at his home. Thus B is chosen in a residential area. In case of a taxicab company, the vehicle is expected to be placed at the head office if it is not driving around. Candidates for these specific locations are not registered as POIs reliably. Hence, they are simulated inside commercial or industrial areas.

- $\boldsymbol{S}$: Like D, the list of approached destinations strongly depends on P as well as on n and the simulated trip types. According to the latter, the target points have to be chosen from all types of POIs and ROIs since taxis travel between various locations. Common examples are the home locations of passengers, stations and airports and additionally detached taxi stands where drivers wait for their next customers.

Compared to the usage models for light-duty commercial vehicles and private passenger cars, the components of IV have more variation. For some of them the size cannot be predetermined but grows during simulation since it depends on the randomly selected trip purposes. No central location from which distances are measured is suitable here. B still is a basis, but it is only approached

to change drivers or at the end of day. All other target points have to be selected subsequently, a parallel processing is only possible for different shifts and days.

3.3.2. Simplification of model

Taxis perform a lot of different single routes. Typical trip purposes depend not only on day and time but also on the geographical region for instance. In some countries taxis are applied in public transport and replace buses on lines that are only used by few people. In Germany for instance, so called hailed shared taxi are common in rural areas. Additionally, taxis are used as patient transport ambulance driving immobile persons to doctor's offices or hospitals and bring them back home after the appointment.[39] In other regions taxis replace school buses or are daily used by commuters on their way to work. [50] for instance splits the trips performed with taxis in 15 groups and compares the partitions for eight cities including New York, Paris and Berlin. These groups do not coincide with the seven categories reported in [38] or [39].
In our reduced model, this large amount of trip purposes is summarized to more general categories, resulting in the following eight different classes.

- **Toward public transport**: The classical taxi trip where people are collected at home or at some other specific location and are driven to a place of public transport like a station or airport.
- **From public transport**: On the contrary to the trip before, people are collected at such a transport nodal point and are taken home. Here, also official visits are included where people are chauffeured to meetings, hotels, etc.
- **Health care**: A trip where immobile people are driven to a doctor's office or hospital. They are collected from home and are brought back after the appointment. These patient transport ambulance drives usually are called on by elder persons or patients requiring some therapy not allowing them to drive themselves.
- **Toward working place**: Especially in huge cities it is common to travel to work by taxi.
- **From working place**: Often used in combination with the trip purpose before, people also take a taxi to return back home after work.
- **From leisure activities and shopping trips**: This summarized group of trips contains rides bringing people back from all kinds of private activities. The take on places include restaurants and bars, shopping facilities, cinemas, sports centers, etc. The destination of the route is always the home of the passenger.

- **Visits and official trips**: With this category, different other kinds of private or official trips are integrated into the model. People can be picked up from any type of ROI and are driven to another. The probability for the classification of ROIs depends on the specific day. During work week business trips are more supposable, thus industrial or commercial areas are rated higher. On weekends, people are expected to conduct rather private visits, hence residential areas are more probable.
- **Return to taxi stand**: This trip type adds the situation that a taxi has no further drive directly after the last one is finished. Instead, a taxi stand is approached where either the next passenger asks for a ride or the central dispatch office announces a new order.

The simplified model does not implicitly include the situation that passengers stop a taxi directly from the roadside. This common procedure usually observable in large cities is not completely neglected. In the evaluation of routes later on it does not matter why the taxi stopped at a specific location. It is irrelevant if the taxi was called there or if it passed by chance and stopped unscheduled. The driven road segments stay the same. Only in case that speed profiles are simulated, an unexpected and a planned stop might cause different results, depending on the base speed of the vehicle.

3.3.3. Returning to the legal driving area

One important constraint for the taxi simulation that influences MP and IV is the legal driving area. Its consideration in the model shall be broached to the issue next, before both parameter sets are determined.
The legal driving or mandatory coverage area indicates the region in that a taxi is licensed for passenger transport. In Germany for instance, drivers have to accept rides inside that zone even if they are short and probably unprofitable. Additionally, special payment guidelines are statutory. In the simulation, the driver should be prevented from departing away too far from this area. Otherwise, a reasonable way back at the end of his shift cannot be guaranteed. Therefore, some correction of the next initial point is performed. Supposed a passenger takes a long ride leaving the legal driving area. Then, after dropping him, the cabdriver starts his way back to his zone, waiting for the next customer. If the next passenger would be picked up near the last one and if that person also wanted a long trip in a direction opposite to the mandatory zone, the taxi would move even farther away. At the end of his shift, the driver might be located more than hundred kilometers away from the place he lives or where he has to hand over the vehicle. Depending on the considered country, he could even not be allowed to pick up passengers that far off his legal zone.
The simulation prevents this problem by projecting back the current position. If the distance between initial starting point and destination of the last passenger is

larger than some given bound, the search for new customers is centered in a point in-between with an acceptable distance. Hence, each time a new trip is started, some "pulling back" is performed first. Figure 3.4 sketches the procedure. As a simplification the legal driving area is assumed to be circular even if this is not correct for most *regions*. Two different distances are employed in order to make the simulation more stable. The larger radius is used to check if the taxi departed too far, the smaller one is applied to determine the next start. B indicates the base fixed in IV. Point A lies near this center and is accepted directly. C is located in a zone which is assumed to be critical. Points lying here are still taken over even if they are outside the boundary. In doing so, not too much positions are corrected. D is settled too far away to be ignored. It is projected back to D'. Of course a smaller translation only towards the outer limit would be sufficient to accept it. However, by this the probability of the adjustment of the subsequent starting point is reduced since the taxicab already drove nearer to the center of the legal zone.

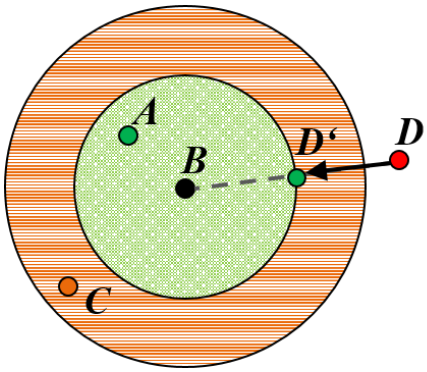


Figure 3.4.: Determination of feasible initial point for passenger search.

3.3.4. Description of MP

The model parameters MP are similar to those of the commuter model. Again, distance distributions for all types of trips are required. Here, a distinction between work week and weekend is appropriate for the frequencies of the trip purposes. Different data sources provide adequate values but nearly none of them include all information needed in an acceptable level of detail. The MiD tables, compare chapter 4.3.1, already mentioned in the description of the commuter model summarize primary purpose, length and days when conducted for trips performed by taxi but do not combine the separate categories. Only the public use file allows a more detailed analysis. However, the match between trip purposes is costly and it is not clear in advance if enough ways are conducted for each case. [50] compares the frequencies of trip purposes for different cities without mentioning driven distances or time indicators. The same holds for [39]. In the

latter some graphics for night rides are given, but those are too vague to be used. More information on the distance distribution is provided in the KiD data [26]. There, distance classes and according frequencies are given split in workweek and weekend. The same holds for the mean number of trips for mobile vehicles. The problem of this data source lies in the summary by trip type. The tables freely available split up the way reported by industrial sector. Taxi companies form no individual category there but are included in class "communications and information transmission". If again the public use file is at hand, a concentration on trips indicated as "collecting, taking and conveyance of passengers" can be identified, but the purpose of the passengers is lost.
MP also has to include the two boundaries introduced in the last section used to keep the taxi near his legal driving zone. Additionally, such radii also have to be provided for the search of the new passengers around this last position. Depending on the statistics at hand, it is not possible to estimate the distance distribution for ways between subsequent passengers. Thus, when one trip is finished, the next customer is looked for in a certain circle around the probably adjusted last stop. If no feasible starting point for the requested new trip type can be found, this circle is enlarged. Only if still no suitable location can be found, the complete route is recalculated. Again, this two stage process is applied in order to prefer starting points in vicinity but prevent the simulation to restart too often. Instead slightly larger distances are accepted. An example for the situation described is the following. Shortly before the end of the shift a passenger ordered a trip to some industrial area. The next and even last trip purpose selected is that of picking up the successive passenger at a transport node and driving him home. From the industrial area, the nearest airport for instance lies in a distance of 25km. If then the lower limit for the search was strict and smaller than 25km, the taxi driver would refuse this request. If some larger upper bound, 30km for instance, was included, the circumstances would be checked and the job would be accepted if the home lies in the legal driving area. The taxi has to drive back anyway and taking a passenger additionally gives some money. The shift can be completed without problems. In the first case with the strict limit, the route gets infeasible and has to be recalculated. Summarized over a huge number of taxis to be simulated, the effort of the re-computation cannot be neglected.

Default values

As already mentioned before, none of the statistics at hand provide all required distance and trip purpose distributions. Hence, a mixture of them is used in the default values. For all trip types the distance classes reported in the tables of [26] are applied. The necessary distinction between workweek and weekend is included there. Like for the commuter model, the distance classes have to be translated from driven to linear ones by use of detour factors, see chapter 5.2.1 and especially section 5.2.2. The lower bounds for the search of new passengers as well as for the

directly accepted last stops are set to 20km. The outer radii are limited to 30km. Here, a simulation for complete Germany including rural regions is assumed, for single cities these values of course might be too large.
The frequencies of trip purposes is obtained by combining [38], [39] and [50]. They are summarized in table 3.1. It also includes the specification of points that have to be simulated later on when setting up S.

Type of trip	Required points	Frequency in work week	Frequency on weekend
Health care	Domicile of patient, POI like hospital, medical practice, ...	8%	0.02%
Toward public transport	Point within all types of ROI, POI like airport, station, ...	20%	12%
From public transport	Point within all types of ROI, POI like airport, station, ...	15%	8%
Toward working place	Home location and workplace of passenger	5%	2.49%
From working place	Home location and workplace of passenger	5%	2.49%
From leisure activities and shopping trips	POI like cinema, bar or supermarket, home location of passenger	17%	60%
Visits and official trips	Point within all types of ROI	15%	10%
Return to taxi stand	Taxi stand	15%	5%

Table 3.1.: Estimated distribution of taxi trips and according points to be simulated

3.3.5. Determination of IV

Setting up IV for the taxi model differs from that of commuters with respect to the order the input variables are created. User types for commercial passenger cars other than taxis are not regarded here, thus U takes only one possible value.

For the pattern P, single and multi-shift operation are randomly chosen. It is assumed that they are of same importance if no further information is given. In contrast to the usage models for commuters and light-duty commercial vehicles, the number of trips n is more important for taxis and influences the size of other parameters. It has to be selected before D and S can be assembled. For each day of the week independent values have to be generated. They are randomly selected from a binomial distribution where the mean and maximal number of trips performed are used to estimate the distribution parameters. In that process, workweek and weekend are separated since usually more trips are performed on the latter. Additionally, if $P =$ *"multi-shift"*, for each shift an individual count is needed.
Afterwards, appropriate numbers of trip purposes have to be selected depending on the frequencies given in MP. These $\sum_{i,j} n_{i,j}$ categories might influence the assembling of the entries of D. It also contains $\sum_{i,j} n_{i,j}$ different distance classes that have to be selected in one of the following ways. If the distance distribution in MP is summarized for all rides, like it is the case for the default values introduced before, D can be filled with a correspondent random vector directly. If individual distributions for different trip purposes are reported, these additional categories have to be respected and more effort is required.
The creation of specific coordinates has then to be performed subsequently. First of all, B has to be determined. It is influenced by P. If a single shift taxi was chosen, B is assumed to be located at the home of the driver. Thus, it is selected inside a residential area. If $P =$ *"multi-shift"*, the base of the car is searched in an industrial or commercial area since it is assumed to be parked at the head office of the taxi company when it is not mobile. These are expected to be settled at central locations with good traffic connections.
After B is fixed, S is filled step by step. For each day and shift the corresponding n_i different trips have to be set up. For this the procedure sketched in figure 3.5 is repeated. As a start, it is necessary to make the last stop feasible. Of course, at the beginning B is accepted directly but for the next iterations this step is essential. For each trip, the location where the passenger enters the taxi has to be simulated first. Figure 3.5 includes a verbal description of the place, table 3.1 already included the classification of the data from which the coordinates are chosen. Afterwards, depending on the chosen trip purpose, the target location of the passenger has to be selected. In case of a drive to a health center, the passenger additionally is returned to his origin. The last stop of the trip is used as the new initial point for the search of the next passenger and is probably projected back to the legal driving zone. However, only the original position is stored in S. In this process, the choice of passenger respects the two radii given in MP, his destination is selected according to the distance class simulated in D. The adjustment of the last points uses the second set of radii.
Only the procedure for trip purpose "return to taxi stand" is slightly different. Here, no passenger is picked up but the taxicab approaches a parking position in order to wait for his next job. Therefore the class given in D is used directly.

All days and shifts are simulated this way independently of each other. Only

Figure 3.5.: Simulation steps in trip generation for taxis.

in S it has to be indicated where they belong to, and their order must not be changed.

3.3.6. Setting up R

The assembling of R is easy now since the driving schedule already was created in IV. For each day, only B has to be inserted between the coordinates given in S. Every shift starts at the base, then all trips are performed and at least B is appended again. In case of multi-shift operation, the single slots are joined and the double entry of B in-between is removed once. In order to tell the different days apart, seven single routes are created for each taxicab. Equation (3.1) summarizes this assembling of R. Index i represents the day, $S_{i,j,k}$ gives all points on the kth trip on the jth shift of day i.

$$R_i = \begin{cases} \left(B,\ S_{i,1,1}, \ldots, S_{i(1,n_{i,1}}, B\right)^T \text{ if } P = \text{"single shift"} \\ \left(B,\ S_{i,1,1}, \ldots, S_{i,1,n_{i,1}}, B, S_{i,2,1}, \ldots, S_{i,2,n_{i,2}}, B\right)^T \text{ if } P = \text{"multi-shift"}. \end{cases} \tag{3.1}$$

4. Working with different types of data

4.1. Geo-referenced data

In order to obtain authentic journeys, all customer simulations should be based on real geographical data. More precisely, the resulting routes should consist of points represented by their lateral and longitudinal coordinates. This allows their projection on maps and thus the routing between them to determine genuine trips respecting the road network in the considered region. We have seen that the different usage models need a classification of target points. Thus the data has to be available in a manageable way including positional and land use information. In the following, a short description of geo-referenced data incorporated in the VMC® database is given. It is based on data provided by the Open Street Map consortium (OSM) [9] and well suited for our purposes. Unless otherwise stated, all geographical data used in the simulation is taken from this source. Nevertheless, an additional different data source, POIPlaza [13], is introduced. It is not the best choice, since it is less complete, but might give an opportunity to enlarge the data basis.
Adapted to the structure of VMC® data, based on the one used by OSM, mathematical methods to select representative target points were developed. These are not limited on this special type of data but can easily be used for other data types using geographical coordinates. Since OSM data has a good coverage, we restrict our considerations and simulations only on this source of information.

4.1.1. Interesting points and regions: POIs and ROIs

In chapters 2 and 3 we have already seen that the usage modeling needs two different types of origins and targets. On one hand, there are precise locations identifying special facilities like supermarkets or hospitals for instance. These locations, commonly called points of interest (POIs), directly consist of latitudinal and longitudinal coordinates mostly representing the middle of the considered building, i.e. a single POI marks a point in S_2 . If a larger structure can be split in smaller parts, like a commercial center for example comprising various stores,

each of the single units should be labeled individually. This enables the distinction between unequal types of classifications which might have different importance for the considered customer group. Furthermore, this finer division allows a better estimation of the density of POIs.
The second type of places are areas for variable kinds of land-use. In the style of POIs, these structures are called regions of interest, ROIs, in the following. They consist of boundaries of polygons with a unique utilization like residential neighborhoods or industrial zones. A single ROI forms a compact subset of S_2.
POIs and ROIs are equipped with markers. These are given as texts and can store various characteristics of the point or area. The general class "sports center" for instance is combined with the type of sports that is exercised there. A "shop" is described further by the category of goods that are sold like food or clothing. OSM applies a system of *keys* and *values* where the *keys* can be used for a generous preselection and the *values* can be taken to select further specifications.
Both types of data can easily be extracted from the VMC® database on a regional or national level. Depending on the usage model of interest, their scope varies and a reduction to relevant units is conducted. An introduction to both types of data and how they can be used is already given in [21, 22]. Here, we want to have a closer look on the data format and how the choice of concrete representative locations can be realized in detail. In the next sections we assume that a selection of relevant data already has taken place.

4.1.2. Simulating with POIs

Points of interest are the simpler of the two data types. We do not consider the size of a building or the volume of sales of a shop, but only its position and category. Due to their limitation on this two pieces of information, simulating destinations from POIs is rather easy. We begin with the situation that the requested POI is the starting point of a trip chain. Thus, there exists only the condition that it has to belong to a specific category or lie in a predefined domain. Hence, we first filter all available points of interest to match these requirements. This step needs no extensive computations but can be done with standard GIS-tools. Afterwards, we list all remaining POIs. Since they all should have the same selection probability, we use a simple random number generator to determine an integer i between one and the number of entities. The point on the ith position in the list is then taken as origin. Figure 4.1 visualizes the procedure on the example of Kaiserslautern as field of activity. The picture on the left shows all regarded POIs inside the city boundary. They have already been colored with respect to their *keys*. The right graphic only contains a reduction to some specific categories of interest. Additionally, one possible choice is marked. The small house indicates that the base position B of the vehicle has been selected for which no distance limits have to be fulfilled.

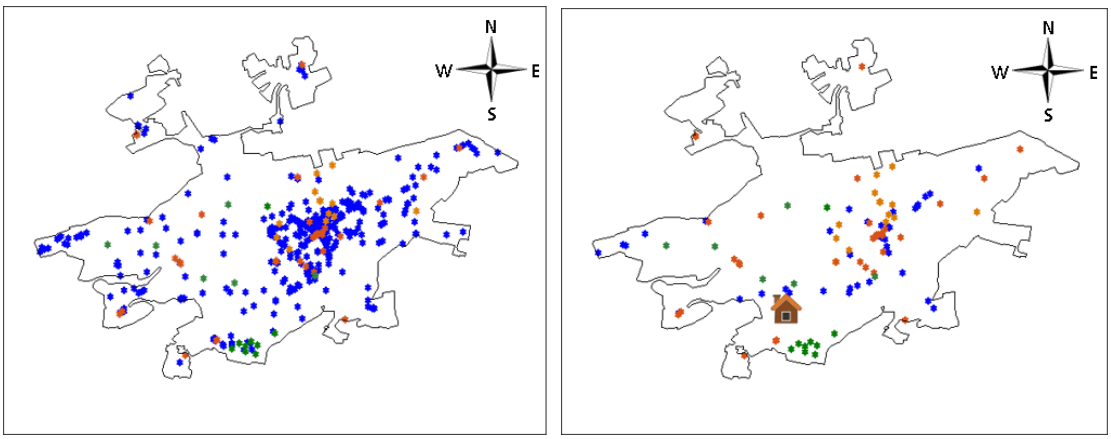


Figure 4.1.: Possible selection of a POI; own illustration embedding [27].

The selection of stopovers or destinations stored in S is done in a similar way. Here, we do not have the claim that the resulting point has to lie in a specific region, but we have some demand on its distance to its predecessor. This distance always has an upper bound, depending on the usage model considered, it might also have a lower one. Again, we use GIS-tools to detect those POIs that fulfill these demands and choose one with help of a random number generator. Hence, we obtain suitable target points in the desired distance. Figure 4.2 depicts the method. On the left, only an upper bound is specified. A circle is drawn around the origin and all POIs not complying the distance limits are shown grayed out. On the right, an additional lower bound is postulated. Thus the number of candidates is reduced. In both pictures one possible simulation result is highlighted.

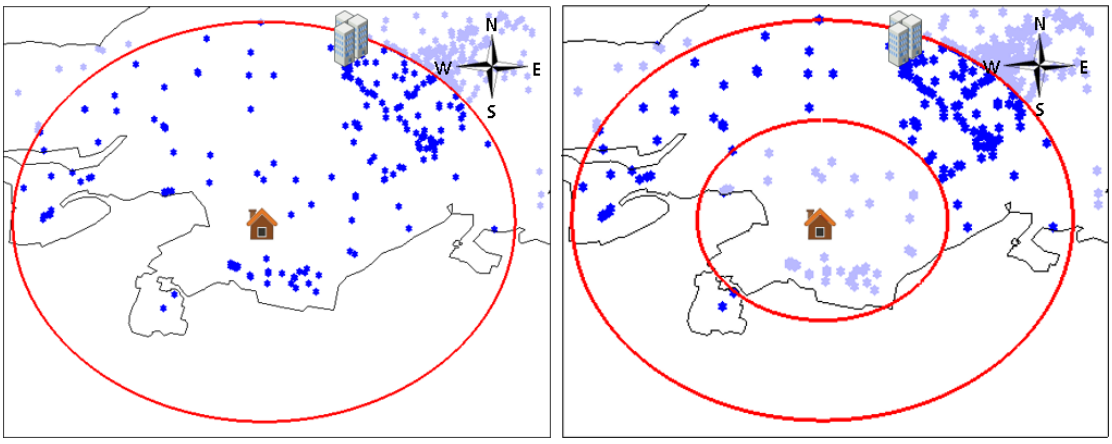


Figure 4.2.: Possible selection of a POI with bounds; own illustration embedding [27, 28].

If the list of feasible candidates is empty, there is no trip possible with the given parameters. Then some simulation steps have to be repeated. When applying the

random number generator, we always assume that all POIs are of same importance and a uniform distribution is sufficient to pick an index in the list. This makes the simulation quite easy but raises the question if this distribution is correct. The answer here depends on the point of view. If the interest lies on the geographical distribution of the simulated destinations, the methods surely provide good results because they were constructed with this intent. We assumed that accumulated POIs are rather attractive because multiple errands can be performed at once. Shopping malls for instance gain greater importance automatically since they consist of a larger number of shops accumulated at one place. Usually, it does not matter which store was the principal destination since ways inside the building are covered by foot and do not influence the simulation results for the vehicle usage.
If the goal is to reflect the regional importance of different types of shops for instance, this simulation surely achieves non-satisfying distributions. In this situation, some indicators of relevance have to be provided. Then we can switch to the cumulative-size method described in [44, p. 225] for example. Alternatively, huge premises associated with just one category which are of larger importance, like universities or hospitals, can be included as new class of regions of interest relevant for special usage models. In the simulation of commuter patterns for example, a category "work" is added. Furthermore, superstores are included in regions of interest classified as "retail" and thus supplementary get importance depending on their areas in comparison to small shops. Therefore we omit the potentially time-consuming retrieval and inclusion of sales statistics or further weighting of POIs in the simulation.

4.1.3. Simulating with ROIs

Regions of interest represent connected parts on the surface featuring the same land-use. Multiple neighboring residential estates for instance are abstracted to one single zone including not only the precise location of buildings but also yards and internal streets. This concept is illustrated on the left part of figure 4.3 for a residential area in Kaiserslautern. Obviously, major streets bordering the zone are excluded but smaller streets inside are not cut out. The screenshot on the right shows a second type of complexity.
Here, not only outer boundaries of the residential area are given, but also the vertices of an inner polygon where a different land-use is tagged. In this case for instance, an allotment is registered. When preparing the regions of interest for simulation, such holes have to be treated carefully. Small embeddings usually do not have an extensive influence on the results. Since the routing algorithm applied later on first projects all sampled locations on the street network, it often makes little difference if the point lies in a small recess. On the contrary, a huge area inside another might falsify its shape and size drastically. This can cause severe errors in the simulation. Thus, the storage format of such *multipolygons* [7] has

to be kept in mind when processing the data. It is described in detail in section 4.1.6.

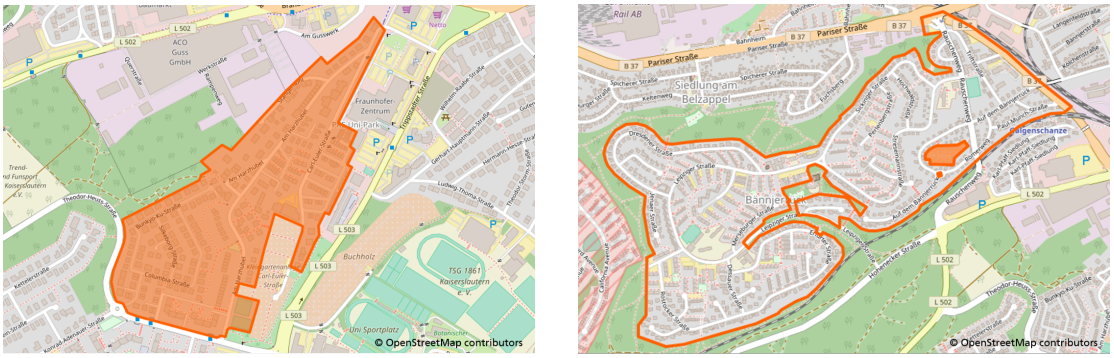
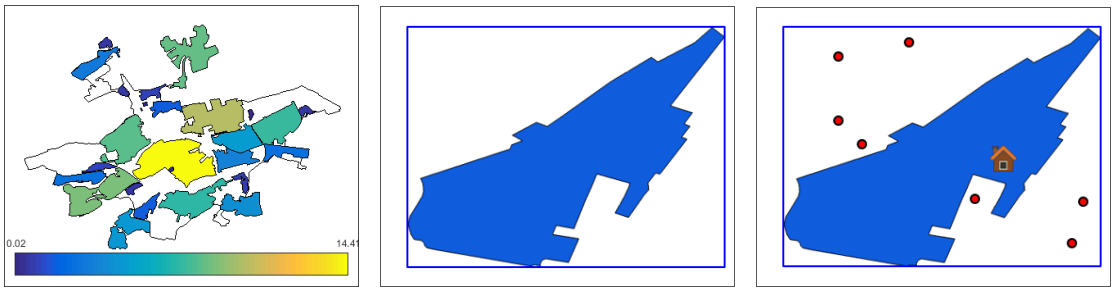


Figure 4.3.: Examples of residential areas in Kaiserslautern. Screenshots are taken from [5]. Left: Simple zone including small streets. Right: Area includes sector that has to be excluded.

The algorithm for selecting representative starting and target points in ROIs requires more computational effort, because here we begin with a heap of polygonal structures and want to simulate coordinates inside one of them. In the common case that there is no further information available on the importance of the single objects, we use the surface area as an indicator for the latter.
When simulating commuters for instance, the population distribution gives a kind of rating to communes and their appropriate residential areas. We postulate here, that this step has already be done, like described in chapter 4.2, and we get a list of ROIs with same relevance. Additionally, we assume that the surface areas have already been calculated. The easiest way to obtain these values is to let them be computed by GIS-tools directly when extracting the ROIs from the VMC® database.
Again, we have to distinguish two different situations. First, we want to consider the case that we start a trip chain and have to find the origin. Then we only have the requirement that the computed point has to belong to one of the given multipolygons. It is expected that larger areas have a larger probability to contain reasonable locations. Thus we take the surface areas as indicators for relevance. Using a cumulative-size method (see [44]), we generate a uniformly distributed random number and trace back to the selected region. Afterwards, we choose a specific origin inside that ROI. We therefore compute the minimum bounding box of the multipolygon and simulate a point inside that rectangle. Finally, we check if this candidate is also contained in the region. If this is the case, the algorithm terminates, otherwise the last simulation step is repeated until a feasible origin is found. Figure 4.4 visualizes this procedure. The first diagram depicts all residential areas in Kaiserslautern and their individual portion on the total surface area.

The second picture shows the selected neighborhood and its bounding box. The third graphic contains some rejected points and the finally chosen one fulfilling the requirements.



(a) All residential areas of Kaiserslautern

(b) One specific neighborhood within its bounding box

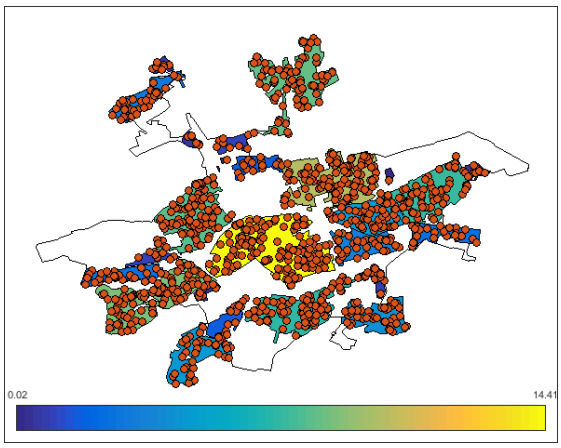
(c) Random selections until success

Figure 4.4.: Selection of a location inside a residential area; own illustration embedding [27].

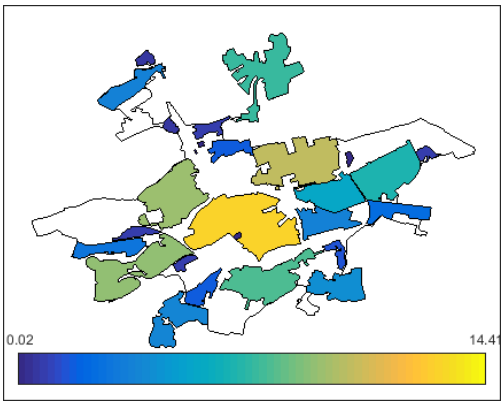
It can easily be seen that the described algorithm produces locations that are uniformly distributed in single ROIs but weighted by area over the complete input data.
Figure 4.5 demonstrates this characteristic on the result of a simulation of 1,000 initial points in residential areas of Kaiserslautern. On the upper, the outcomes are recorded at their simulated location. The second picture visualizes the fraction of obtained homes for each residential area. The last graphic repeats the distribution based on surface areas to facilitate the comparison. Obviously, the selection algorithm produces appropriate results. Small areas are chosen less often than larger ones. The huge zone in the center gets the most importance in both cases. Even the smallest neighborhoods are sometimes selected as base points. The holes are left blank like expected.

In the second situation again a target has to be simulated. In addition to the classification of ROIs it has to lie in, it also has to fulfill some distance condition to its predecessor. Like already seen for POIs, an upper limit always has to be respected. Depending on the trip type, also a lower bound may be given. In a first step, we thus reduce the list of correctly categorized ROIs meeting these conditions. For that purpose, a circle with the upper distance bound as radius is drawn around the predecessor. In the case of an additional lower bound, a circular ring is constructed. Both alternatives can be summarized to one procedure if we allow the minimal distance to be equal to zero. Next, we determine those ROIs that have a non-empty intersection with the circular ring. The result is sketched in figure 4.6.

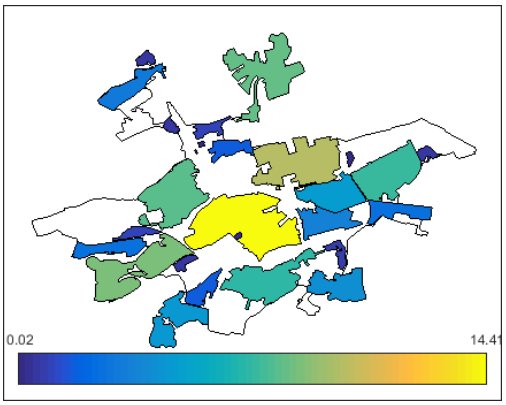
The simulation of a specific point is performed on basis of some kind of polar coordinates. Two uniformly distributed random variables are generated such that



(a) 1000 home locations simulated in Kaiserslautern, colorbar indicates input distribution



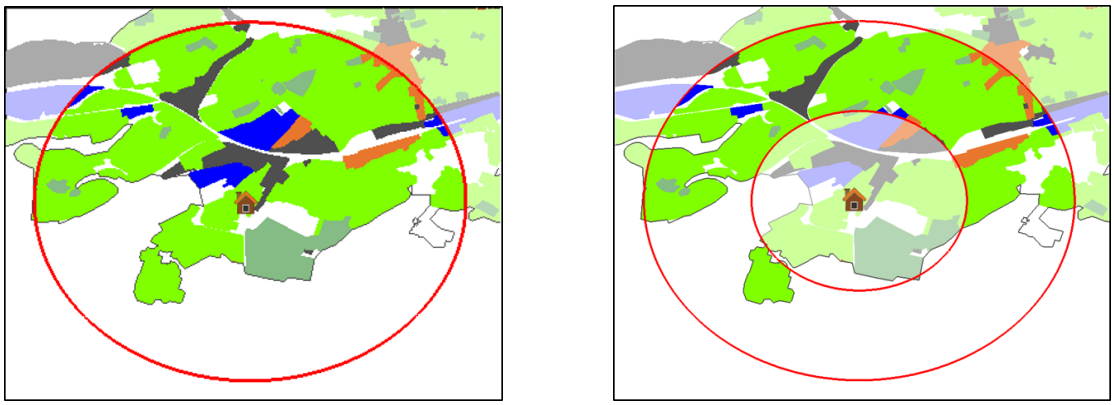
(b) Fraction of simulated locations in areas



(c) Fraction of areas regarding surface area

Figure 4.5.: Result of a simulation of 1000 home locations in Kaiserslautern. The colorbars indicate the fraction in percent.

the first is chosen between the two distance limits and the second has to lie in the half-open interval from zero to 2π. Obviously, the first value represents the chosen distance between origin and target on the surface. The second gives the initial bearing. Afterwards, spherical geometry is applied to calculate geographic coordinates on basis of these parameters using great-circle distances. More information on this topic can be found in appendix A. The resulting point then lies somewhere in the requested circular ring and it has to be decided if it belongs to the reduced amount of ROIs.
If this is not the case, the random number generator is started again. The procedure is repeated until a valid location is found. The algorithm is constructed such that the cross section between ROI and circular ring indicates the probability of a target point being settled in that region. Within all permitted surfaces,



(a) Only an upper bound prescribed

(b) Also a lower bound given

Figure 4.6.: Determination of locations in ROIs in feasible distance; own illustration embedding [27].

the generated points are distributed uniformly. Figure 4.7 sketches this result.

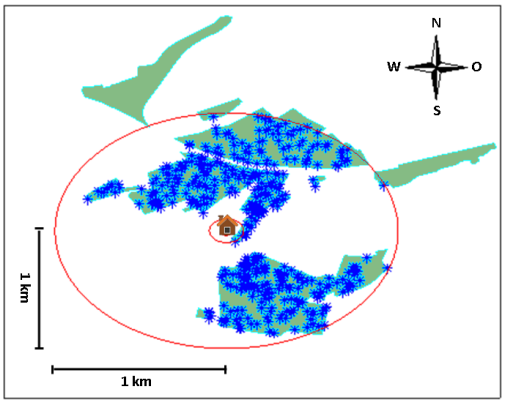


Figure 4.7.: Determination of areas in feasible distance; own illustration embedding [27].

The reduction of all correctly classified ROIs to the ones intersecting the circular ring might seem unnecessary at a first glance, but it serves two purposes. On one hand, it helps reducing the number of checks, if the candidate point is contained in a ROI. Considering the simulation in a large country, the number of comparisons is hence shortened drastically. The more important point is the avoidance of a futile candidate generation. It is possible that the requested distance class does not match the local circumstances and that there is no suitable ROI left. Then the desired trip is not feasible. The treatment of such incidents is discussed in chapter 5.

4.1.4. Simulation of stopovers on given trips

In the last two sections the simulation of initial and target points was described in detail. However, a special situation has not been discussed yet. It deals with the circumstance that the main parts of the trip already have been determined but afterwards, a stopover lying in-between is requested. For the sake of simplicity, the distance classes for both tracks are expected to be available, i.e. the joint distribution of the two rides splitting the original one is known.
The easiest way to include the extra destination in the given trip is to expand the illustrated algorithms by drawing two circular rings instead of only one, like shown in figure 4.8. This postulates that the additionally driven kilometers should not dominate. However, obviously the amount of suitable POIs and ROIs might be very small. Depending on the considered usage model and even the trip purpose, this limitation is acceptable or not. In the case of commuter patterns for instance, it is expected that people avoid too large detours doing shopping trips on their way home from work. Parents dropping off their children at school or kindergarten rather accept moving around because these stopovers are fixed and allow no variation. For them a restriction to a regional based selection instead of a distance based might be better. Alternatively, the one nearest to the place of residence might be a good choice.

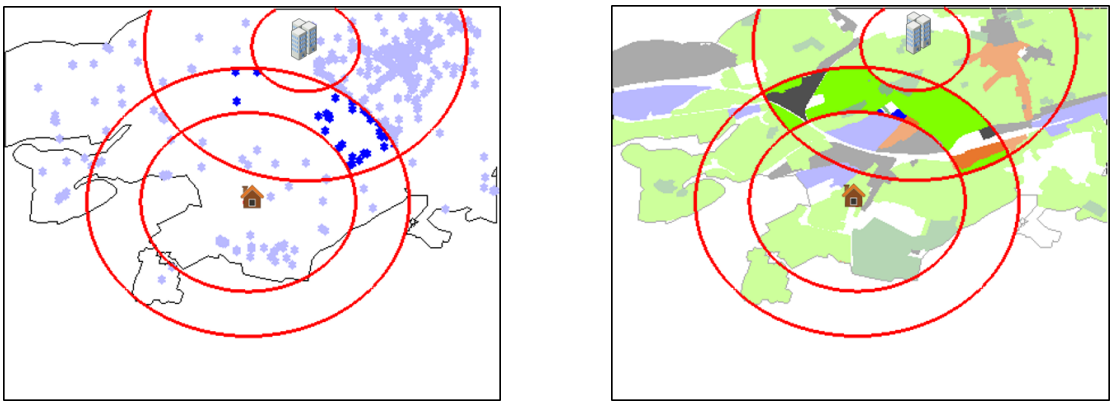


Figure 4.8.: Suitable destinations are reduced drastically when multiple distance classes are demanded; own illustration embedding [27, 28].

These two examples show that no general statement for the handling of intermediate destinations can be made. Depending on usage model and trip type, different conditions influence the decision for or against such a two-sided stopover determination. The algorithms simulating locations from POIs and in ROI presented before can easily be adapted to these requests. The problem in the majority of cases lies in the availability of the joint distribution of the combined distance classes.

4.1.5. Benchmark of OSM data

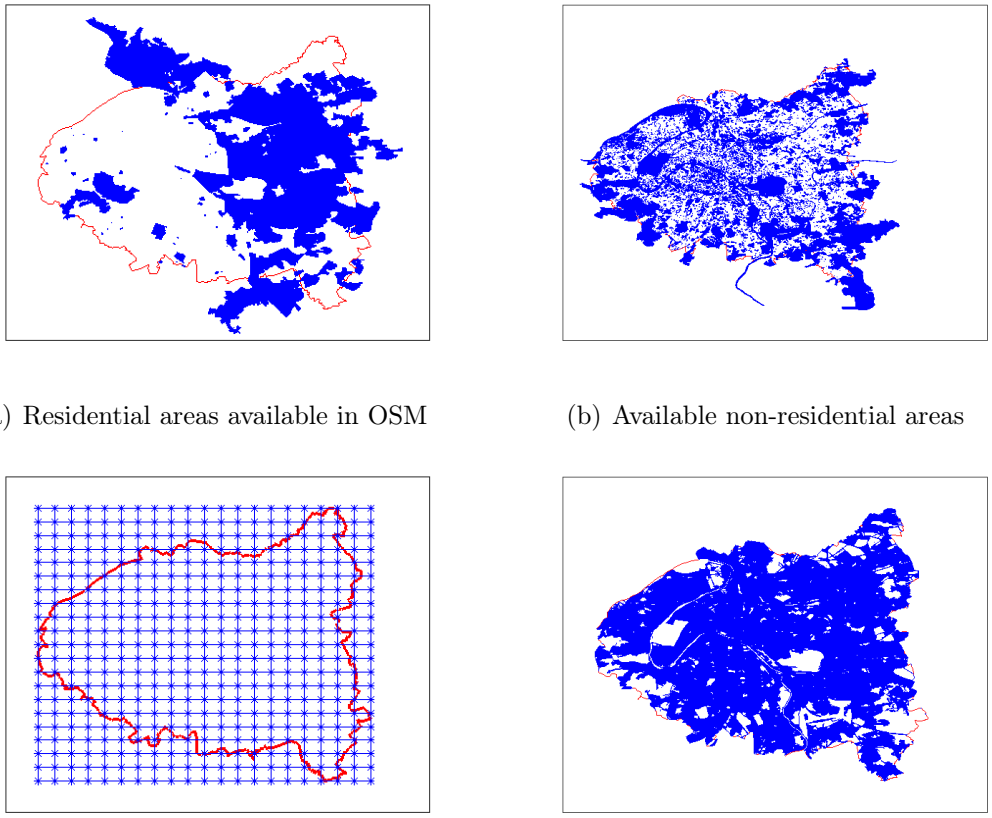
The benefits of using POIs and ROIs from the OpenStreetMap project [9] have already been discussed in detail in [22] and [21]. A summary is given here. The main advantage of this database is its good coverage for most countries of the world in combination with its flexibility. Undoubtedly, the quality of data depends on the considered region and is not as good in isolated areas as in metropolitan areas since it subsists on the participation of the over two million people [11] contributing their collected data. However, errors are fixed continuously and the coverage grows constantly.
In contrast to other sources, OSM contains POI and ROI data. In the majority of the cases, alternative data providers offer POI information in various different categories, but exceedingly few also supply land-use data. However, this is essential for the usage simulation. If ROIs can yet be extracted from other sources, high costs are imposed. OSM data in contrast to that is free of charge and it lives from the work of volunteers, donations and material support [10].
Both data types extracted from OSM, regions and points, can rather easily be preprocessed for the usage modeling. As already mentioned before, the attached markers consisting of *key* and *value* tags can be used directly for a classification of single data points. Since POIs are given by their geographical coordinates, no further work is needed for them. Regions of interest require a bit more effort due to the employed *multipolygons*. Their preparation for non-GIS tools like MATLAB is explained in the last section subsection 4.1.6. However, even this does not take much time and only has to be conducted once for multiple simulations in the same region. Thus, the data format is still easy to handle.

Reconstructing missing neighborhoods

One issue arising using OSM data is the lack of tagged neighborhoods in some cases. Having a look at the city of Paris for example, the map shows a reasonable distribution of residential areas including numerous housing developments, see figure 4.9. When the ROIs for the town center are extracted from the database, only few residential areas are obtained, nowhere near covering the complete district. The available zones are sketched in figure 4.10(a). This difference between map and database is a result of the map renderer filling the complete area inside the city boundaries. Often, "special" land-use like industrial, but also parks or graveyards, are registered and thus drawn on top. The obvious residential areas are neglected and not tagged and contained in the database. Since these are indispensable in the usage models, an auxiliary construction is needed.
In the following method we assume that residential neighborhoods are the only category missing or incomplete. Then it is possible to take the administrative boundary and subtract all fields with a different land-use reported. We result in the most likely residential areas of the town represented as one single structure. In

Figure 4.9.: Screenshot of the map of the city center of Paris taken from [5].

small communes with small areas, this structure can be simple and sufficient, but for larger cities it is expected to be very complicated. The generation of random points and especially the check if a candidate is valid, is then very time-consuming. Hence, a split into smaller pieces is preferred.
Some kind of individual "look and cut" procedure surely gives good and logical solutions, yet it is involving and inefficient. A general algorithm dividing the area automatically is favored. The method applied for Paris for instance computes the bounding box and lays a grid upon it. For every emerging square or rectangle an intersection with the obtained structure is performed. Parts inside the box not including fractions of residential areas are deleted directly. Non-empty intersections are stored as new ROIs in combination with their surface area and classification as residential. After all blocks have been processed, the large structure is deleted and the new regions are added to the list of ROIs used in computation. Additionally, they could also be added directly to the database.
The size of the resulting areas strongly depends on the constructed mesh. Its density has to be adapted to the shape and the size of the considered *region*. A benefit of the approach lies in its flexibility. Squares are not the only possible base, also a triangulation can be performed in advance to the cutting procedure.



(a) Residential areas available in OSM

(b) Available non-residential areas

(c) Grid over Paris

(d) Obtained residential areas

Figure 4.10.: Computation of residential areas in Paris.

Ancillary data sources

After the close look on the OSM data integrated in the VMC® database, we also want to review some other sources. Here, POIPlaza [13] is one prominent free of charge alternative. On this website one can download POIs of various categories for single countries.
Its advantage lies in the manifold of supported data formats. Users can obtain for instance all stores of one supermarket chain or all fuel stations of one operator. The achieved information can then directly be imported to navigation devices. For our purposes it can also be used easily in the usage modeling. However, like the name says, POIPlaza does not provide further data. The same holds for the fee-based alternatives POIbase[12] or GPS Data Team[3].
Since maps are already available on navigation devices and updates are released regularly by the producers, people usually are not interested in ROIs. Land-use information can hardly be found in any other source, least of all low-priced or free. The usage models only can benefit from additional POIs when countries are

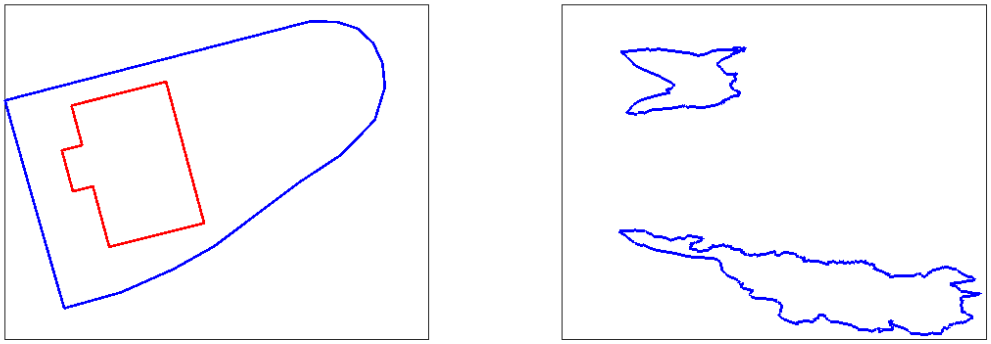
better covered than in OSM or if special sector-specific destinations are required that can be downloaded directly instead of performing a complex search. For our purpose, the VMC® database is of sufficient quality and offers everything needed so we base our simulation only on this available data.

4.1.6. Multipolygons

Multipolygons are the common data type for ROIs as well as administrative borders defining *regions* for the usage simulation. A description and some examples for feasible and invalid structures can be found on [7]. Here, we use areas included in the VMC® database as examples to demonstrate the concept. In addition, we turn our attention to the translation of *multipolygons* for the processing with MATLAB.
As the name indicates, *multipolygons* are a combination of several single polygons. These contain a list of the coordinates of the vertices, see [15]. Like described on [14], polygons are intuitively composed of a necessary outer boundary b_{outer} and possibly inner rings b_{inner}. The latter mark zones that are excluded from the overall shape. In order to explain this concept, we consider a parking area in Kaiserslautern. This type of land-use plays a minor role in the usage modeling up to now, but due to its clear and easy profile it is suited as a demonstration example. Its geometric description extracted from the VMC® database reads

MULTIPOLYGON(((7.7506002 49.4333364,7.7505018 49.4335709,7.7515812 49.4337596,7.7516763 49.4337577,7.751747 49.4337424,7.7518025 49.4337075,7.7518355 49.4336591,7.751844 49.4336028,7.751808 49.4335241,7.7516886 49.4334411,7.7515436 49.4333768,7.7512403 49.4332231,7.7510963 49.4331686,7.7509117 49.4331135,7.7507111 49.4330755,7.7506002 49.4333364),(7.7507368 49.4335589,7.7507747 49.4334639,7.7507028 49.4334518,7.7507417 49.4333544,7.7508125 49.4333663,7.75087 49.433222,7.7512052 49.4332785,7.751071 49.4336153,7.7507368 49.4335589)))

its visualization is given in figure 4.11(a). b_{outer} is painted in blue, b_{inner} is given in red. Obviously, both components can be found in the multipolygon, just split by a comma. Similar, a true multipolygon, that is not just a polygon, can then be obtained by combining several of such structures. Usually, its description then gets rather lengthy and shall be skipped here. Figure 4.11(b) depicts the administrative boundary of the federal state of the Free Hanseatic City of Bremen. It is split in the city municipality of Bremen and the exclave of Bremerhaven. The polygons are not connected geometrically but form one political unit. Multipolygons enable their storage as one entity.



(a) A parking area in Kaiserslautern. (b) The federal state of Bremen.

Figure 4.11.: Sketches of two geometric representations.

In contrast to MATLAB for instance, where such structures are described by the coordinates of the vertices ordered clockwise and counter-clockwise, the single components are connected using parenthesis.

4.2. Estimating vehicle distributions

One main issue in the simulation of vehicle usage is the choice of initial locations. The analysis of driven roads strongly depends on the region people are traveling around. Thus, the distribution of vehicle population has to be reflected well. Since the computation of further destinations uses the home location of the vehicle as a basis for distance calculations, discrepancies in these data might cause huge errors in the final results.
The problem in estimating the correct distribution is the availability and accessibility of data. Surely, for each country of the world some vehicle statistics or at least some population figures exist, but they have to be found and included in the database before the simulation can be started. We want to omit this additional effort and use only data that is already at hand in VMC®. However, we first have to check the quality of this. The next sections summarize how this can be done on the example of Germany. In that process we compare population counts reported in OSM with those from official statistics. Afterwards, we show how the simulation of the usage model handles this data. In the end we also check if additional official vehicle statistics are required or if they can be estimated from data at hand.

4.2.1. Working with population figures

For the comparison of population counts, first a suitable data base has to be chosen. In Germany for instance, population figures are published by the Federal Statistical Office several times a year summarized on different levels. The number of inhabitants of the Federal states is released quarterly in [56], the population counts based on rural districts and urban municipalities are published once a year reporting the numbers of the last year. The data used for the comparison, [55], includes the average population over the year 2015. Since data needs some time to be forwarded to OSM, we take the VMC® database from 2017 as reference. It was extracted in December 2016 and hence might be too close to the release data of the statistics. However, it is the newest available. Additionally, we assume that population figures do not change that much between subsequent years and are not adapted that frequently in OSM. Even some larger differences, like the consolidation of two rural districts in 2016, are not reflected, yet. Thus, a comparison between statistics for 2015 and OSM data from 2016 is admissible.
Unfortunately, the data that shall be matched cannot be employed directly. It is provided on different levels and has to be linked first. Some more or less preprocessing has to precede the actual comparison.

Considering available statistics

Official statistics about population figures like [55] group the counts by administrative districts which are only represented by name and some official identification number like the "Official Municipality Key" in Germany. In some rare cases they also contain the coordinates of the city center, but this is rather an exception than a rule. Usually they lack a geographical position. Some reference between lower and higher level administrative units, like rural districts and federal states, is often only given by the municipality key. For entities of the same level, no neighboring relation can be determined. Nevertheless, statistical data can be processed easily in order to extract corresponding population numbers and municipality keys.

Preparing VMC® data

In the OSM database the population counts are not stored consistently. Unfortunately, these problems are passed to the VMC® database. According to [8], population numbers commonly are reported at nodes marking the center of a commune or other administrative unit. Additionally, they are often adjoined to the description of the boundaries. The same inhomogeneity holds for the municipality keys. Usually, they are provided for those areas but are as well given at nodes,

compare [16] and [17]. Consequently, there exist two sources containing the required information which are not directly linked in OSM. In order to combine them, first a check for inconsistent and double municipality keys resulting from differing authors or update times is conducted. In the worst, and regrettably existing, case, a commune is even not enriched with the required information in any source. Thus, a more geometric match is required.
The rather easy test for nodes lying in the multipolygons describing the administrative boundaries has to be handled with care. Again, different levels of units are available. Thus, the nodes are naturally contained in several of them. They have to be distinguished by the *key-value* pair (compare chapter 4.1.6) marking the levels. Suburbs and rural districts are skipped, only places tagged as *municipality*, *town*, *city* or *village* are considerd. Then, population numbers for larger units can be computed as the sum of the smaller ones lying inside. The check for doubly inserted keys twice performed in advance should prevent that entities are added multiple times. As a result, we obtain a table including all VMC® data needed in the comparison like municipality keys, if available, population figures, computed from smaller units, and geographic coordinates of the boundaries and their administrative level.

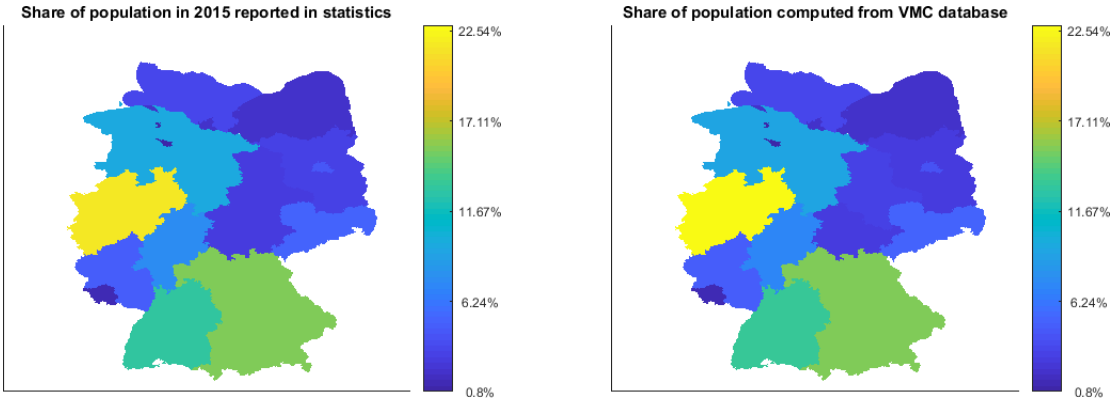
Linking both data sources

Before official statistics and population numbers in VMC® can be compared, they also have to be matched first. The goal is to have both figures for each administrative unit of interest. Here, we restrict ourselves to federal states and rural districts for two reasons. First of all the data provided in OSM on a commune level seems to be defective. For instance, there are municipality keys included that cannot be found in the statistics. Since local government reorganization is performed every year, new keys are created and old ones are removed regularly. Often this is not transferred to OSM directly. In addition, the number of communes is high but not consistent between data extracted in subsequent years. Hence, the extensive check for changes has to be performed every year for every country. However, the main point for the restriction to larger units lies in the availability of statistical data. For Germany for instance, population counts up to the commune level are published. Vehicle statistics in contrast are only provided on a rural district level. Since the overall goal is to reflect the vehicle distribution well, no smaller units can be included in the comparison. Thus, the check for municipalities gives information on the quality of the OSM data but cannot be used later on for the usage modeling.
Consequently we concentrate on the 16 federal states and 402 rural districts. The match between both data sources shall be performed via the municipality keys. Unfortunately these are not reported thoroughly in many cases, even not in the official statistics. Actually, this code consists of different blocks of fixed size making it possible to identify regions of various levels of importance like federal states

or rural districts. However, often they are reduced to the "relevant" digits. We do the same for the comparison. We remove leading an closing zeros and only take the remaining numbers. We only have to guarantee that "10" is kept and not reduced to the already available number "1". Additionally, the city states have to be included as rural districts, also. Afterwards, all administrative units needed are correctly linked and can be compared.

Comparison of official statistics and data given in the VMC® database

After the matching of data, each administrative unit of interest is equipped with several pieces of information. These include the geographic description of the boundary, the name, the reduced municipality key and population counts from statistics and VMC®. The latter are compared in two different ways. First of all and suitable for federal states as well as rural districts and urban municipalities, a map of Germany can be drawn for each data source and level. The absolute population counts here only play a minor role. It is not important if we have differences up to some hundred inhabitants as long as the relation between the units is reflected well. In the simulation of the usage model only a fraction of the population is selected.
Hence, we sum up the counts for the complete country and calculate the fraction of each unit on this. Afterwards, we draw maps including colorbars indicating the proportions. Figure 4.12 shows the result for the federal states, figure 4.13 depicts that or rural districts and urban municipalities. Obviously, the correspondence is quite good. For the federal states we only have slight differences in Saxony-Anhalt.

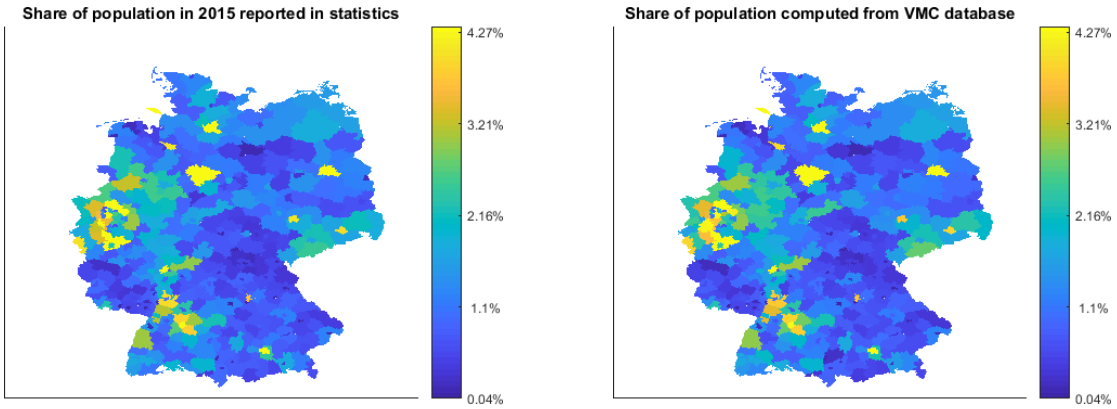


(a) Population distribution reported in [55]

(b) Population distribution computed from VMC® database

Figure 4.12.: Comparison of population distribution for federal states. The proportions of inhabitants are indicated by the coloring.

For the smaller administrative units the result is similar. Some small entities show a minor change in color, but the overall result is acceptable. As a second



(a) Population distribution reported in [55]

(b) Population distribution computed from VMC® database

Figure 4.13.: Comparison of population distribution for rural districts and urban municipalities. The colors illustrate the proportions of inhabitants.

tool for comparison, we directly face the resulting fractions and for the sake of completeness also the absolute numbers. Since this gets confusing for the 402 smaller units, we concentrate on the federal states. Figures 4.14 and 4.15 include both charts. It can be seen that only in few cases the absolute population numbers coincide. For the proportions the situation is a bit better. Here, about half of the fractions have a difference less that 0.1 percentage points.
Hence, at least for Germany, no population figures from official statistics are required to determine the population distribution sufficiently well. The counts included in the VMC® database are of good quality and can be used directly.

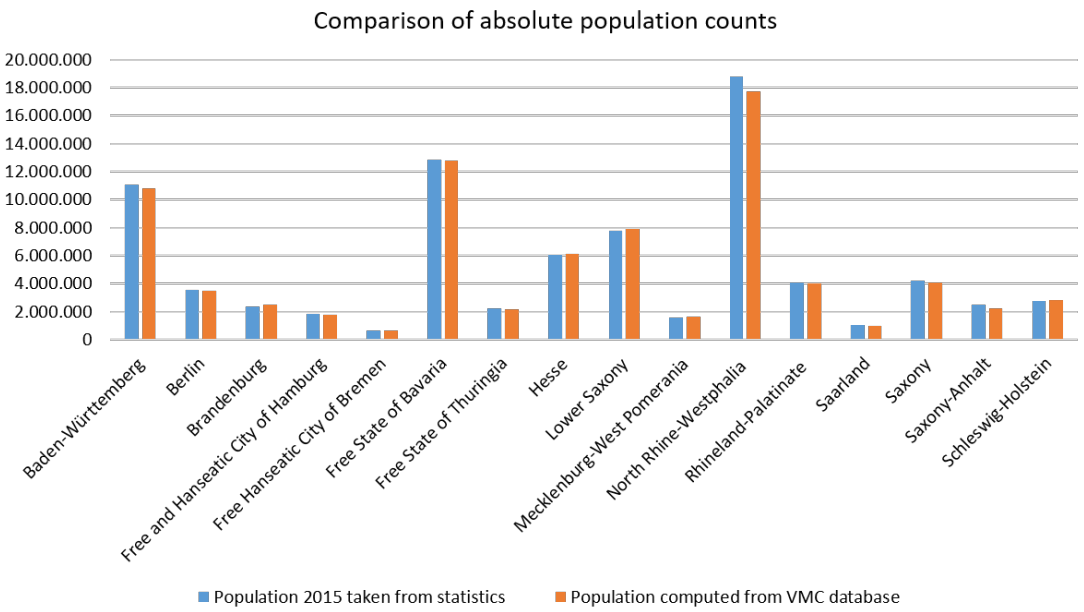


Figure 4.14.: Comparison of absolute population values of statistics for 2015 and VMC® database.

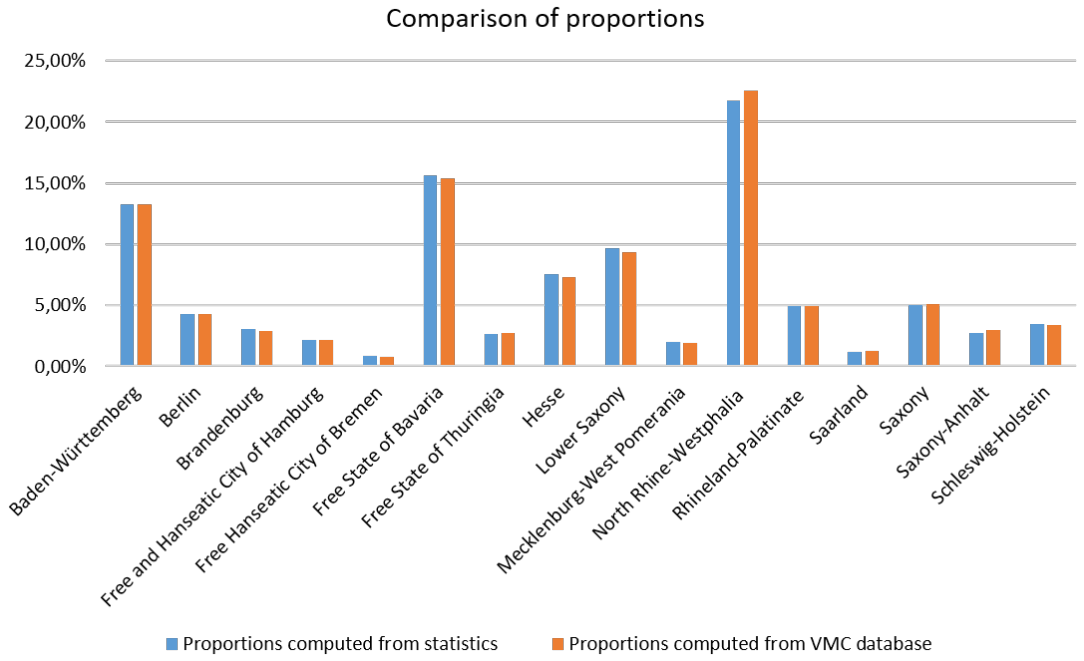


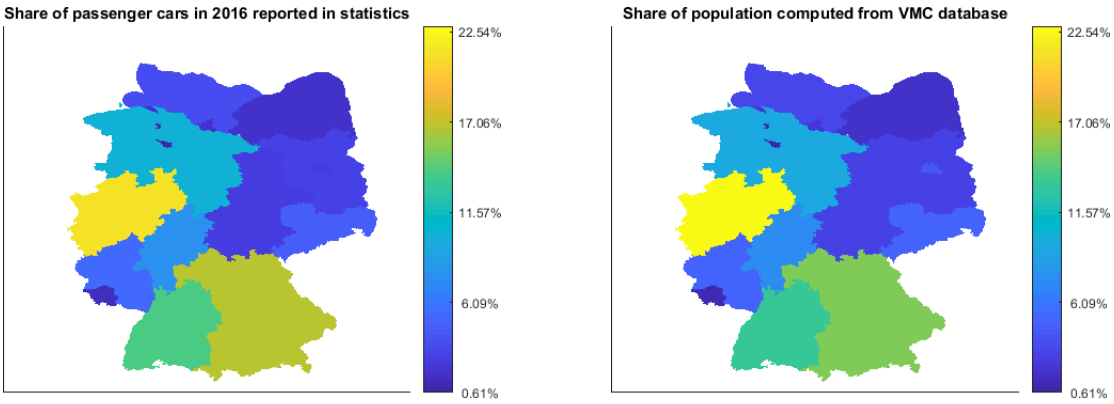
Figure 4.15.: Comparison of population distribution of statistics for 2015 and VMC® database.

4.2.2. Including the vehicle distribution

Up to now we have shown that the true population distribution provided by official statistics is reflected well by population counts included in the VMC® database. In this process we concentrated on federal states and rural districts and urban municipalities. These are sufficient for the comparison with the vehicle distribution. The determination of a good estimator for this is the actual goal of this chapter. As a basis, we use the number of passenger cars in Germany on January 1st in 2016 published in [41] in April 2016. In this report the counts of all kinds of registered vehicles are provided, but for the sake of simplicity we concentrate on this specific group. We also do not distinguish between private and commercial ones. They exhibit similar shares for most of the units with same classification. Again, like for the population figures, the numbers are summarized on the two levels of administrative units already mentioned. There is no information available allowing the assignment to smaller communes like villages.

Comparison of population and vehicle statistics

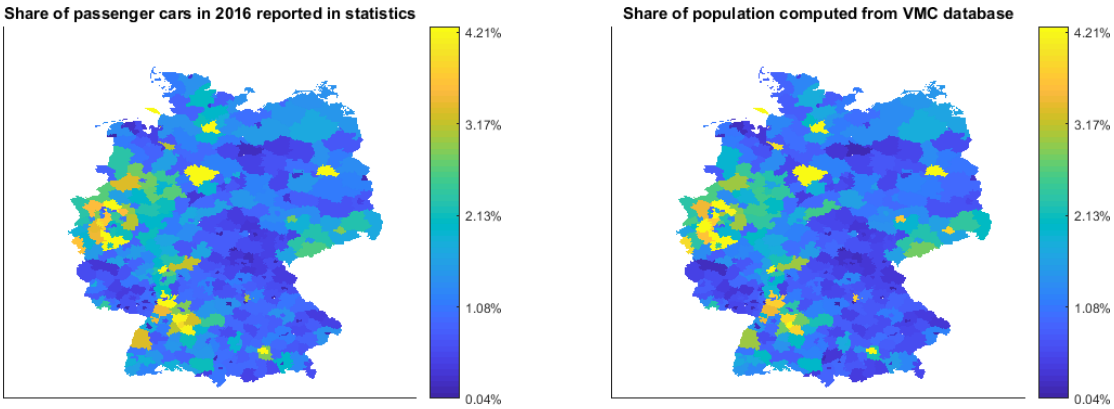
It would be comfortable if the vehicle distribution could be estimated from data already available in the VMC® database. Otherwise, each time a simulation of a usage model is performed, a search for official statistics and their inclusion into the existing structure is necessary. It is assumed that the vehicle distribution strongly depends on the population distribution, thus these shall be compared first. Other candidates for influencing factors are the area of the complete administrative units or that of the residential neighborhoods. Both can be computed from available geographical data. A different parameter could be some further classification of districts depending on their regional importance. However, this specification of so called *Kreisregionstypen* is special for Germany. Since it is not intuitive and several entities have to be combined, it cannot be transfered to other countries. Thus, it is not suitable as a general factor.
Figures 4.16 and 4.17 sketch the comparison between vehicle and computed population distribution. For most of the federal states the results look quite well. Again, there is only some slight overestimation for Berlin, North Rhine-Westphalia and Saxony-Anhalt and an underestimation for the Free State of Bavaria. In the match for the smaller units, the Ortenaukreis and the region around Stuttgart have a smaller proportion of inhabitants than passenger cars. For Leipzig and Dresden it is the other way around.
For a more detailed analysis we concentrate on the federal states. We have a look at the precise numbers and try to enhance them by using the available parameters already mentioned, namely the two types of areas.



(a) Passenger cars reported in [41].

(b) Population distribution computed from VMC® database

Figure 4.16.: Comparison of passenger cars reported in statistics and computed population distribution for federal states. The colors illustrate the proportions for both quantities on their overall sum.



(a) Passenger cars reported in [41].

(b) Population distribution computed from VMC® database

Figure 4.17.: Comparison of passenger cars reported in statistics and computed population distribution for rural districts and urban municipalities. The colors illustrate the proportions for both quantities on their overall sum.

In figure 4.18 the single shares of the available quantities are opposed. For each factor, its overall sum is computed and then the proportion of each administrative unit is calculated. It can be seen that for these initial values the population distribution resembles most that for the passenger cars. The areas seem to behave differently compared to the vehicle counts. However, we are interested in the magnitude of the error we make when using the population distribution only and its reduction due to inclusion of additional factors. Thus, we compute the expected difference in the shares. For the sake of readability we assume the population

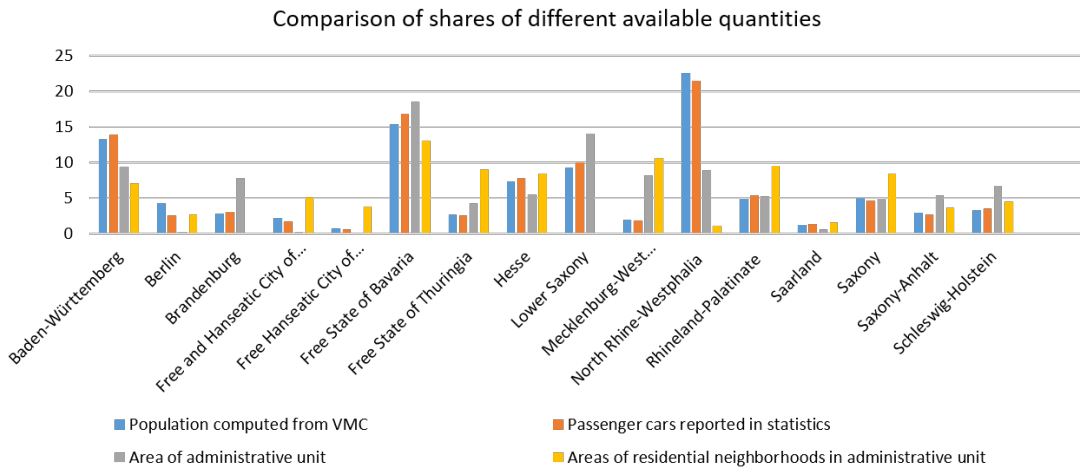


Figure 4.18.: Comparison of proportions for available quantities.

figure to be completely described by a vector $x_1 = (x_{1,1}, ...x_{1,16})^T$ summarizing the proportions for all federal states. Similar to that the distribution of passenger cars is given as $x_2 = (x_{2,1}, ...x_{2,16})^T$. So the first index of x distinguishes the two distributions, the second represents the federal state. Recall that $x_{1,j}$ is the proportion of population coming from state j. In particular, $x_{1,1} + ... + x_{1,16} = 1$. The same holds for passenger cars and x_2 analogously. If we want to pick a car in a representative manner from the whole population, we would choose the state from a multinomial random vector $\eta = (\eta_1, ..., \eta_{16})$ with parameter $(1, x_{2,1}, ..., x_{2,16})$, i.e. one $\eta_j = 1$ and all other η_i, $i \neq j$, are 0. The expectations are $E\eta_j = x_{2,j}$, as the individual η_j are Bernoulli variables.
If x_2 is not known, we might use x_1 as an approximation, resulting in a corresponding multinomial η^* with parameter $(1, x_{1,1}, ..., x_{1,16})$. The expected error is

$$E\left(\eta_j^* - \eta_j\right) = x_{1,j} - x_{2,j}, \ j = 1, ...16. \tag{4.1}$$

However, we are more interested in the relative error which we define as

$$\frac{E\left(\eta_j^* - \eta_j\right)}{E\eta_j} = \frac{x_{1,j} - x_{2,j}}{x_{2,j}}, \ j = 1, ...16. \tag{4.2}$$

This has the advantage that it does not depend on the number of cars used in the simulation. If we would generate z cars, then the numbers η_j from state j form a multinomial random vector with parameters $(z, x_{2,1}, ...x_{2,16})$, and correspondingly for η_j^*. The expectation is $E\eta_j = x_{2,j}z$ respectively $E\eta_j^* = x_{1,j}z$, and the relative error

$$\begin{aligned} \frac{E\left(\eta_j^* - \eta_j\right)}{E\eta_j} &= \frac{x_{1,j}z - x_{2,j}z}{x_{2,j}z} \\ &= \frac{x_{i,j} - x_{2,j}}{x_{2,j}}, \quad j = 1:16 \end{aligned} \tag{4.3}$$

independent of z. An overview for all federal states is given in figure 4.19 in the blue bars. Obviously, for most of them the error is small but there are three outliers, Berlin, Hamburg and Bremen. Hence, the city states seem to behave differently than all others, which does not come as a surprise. This dissimilarity is also reflected when the ratio of residential and complete area is computed, see figure 4.20. These three states show a huge ratio whereas the others' are bounded by a ratio of less than 1%.

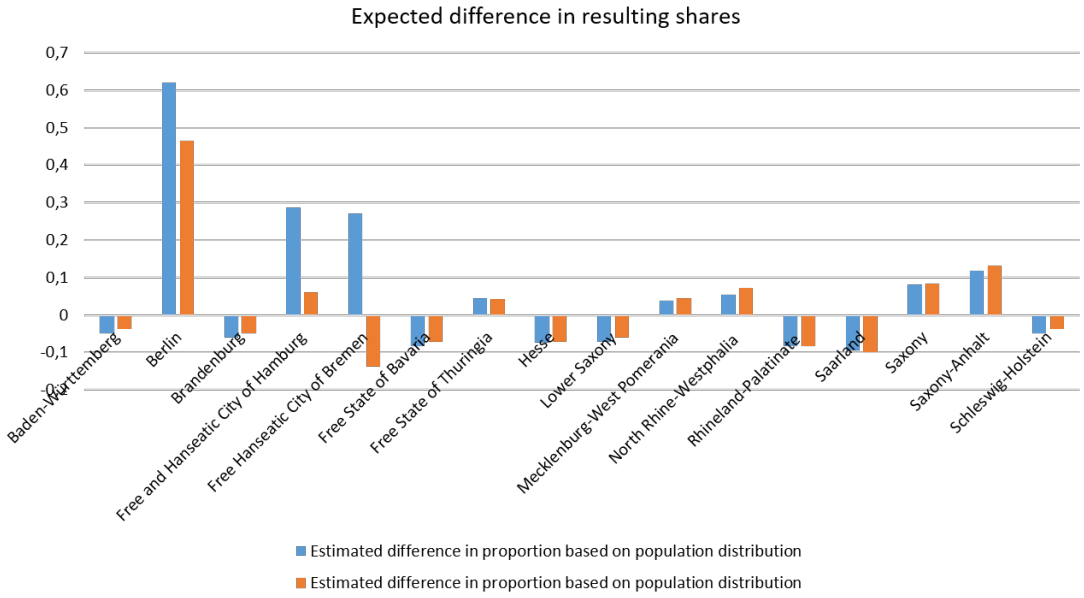


Figure 4.19.: Comparison of expected difference in shares after simulation.

As a remedy, x_1 is replaced by some modified vector x_1' of multinomial probabilities, such that

$$\frac{x'_{1,j} - x_{2,j}}{x_{2,j}} \approx 0 \tag{4.4}$$

for all $j = 1, ..., 16$. Therefor some entries of x_1 have to be reduced by a certain amount to be calculated. We proceed in the following way:
First of all we decide that the correction shall be performed by a multiplication with some factor, i.e. $x'_{1,j} = p_j x_{1,j}$ with $p = (p_1, ...p_{16})$ depending on the ratios

$$t_j = \frac{A_{res,j}}{A_{complete,j}}, \ j = 1, ..., 16, \tag{4.5}$$

of residential and complete area of administrative unit j. This ratio is large for urban areas, here in particular for the three city states, and small for rural areas. Regarding (4.4) with equality, the ideal choice would be $x'_{1,j} = p_j x_{1,j} = x_{2,j}$, i.e. $p_j = \frac{x_{2,j}}{x_{1,j}} = p_j^0$.

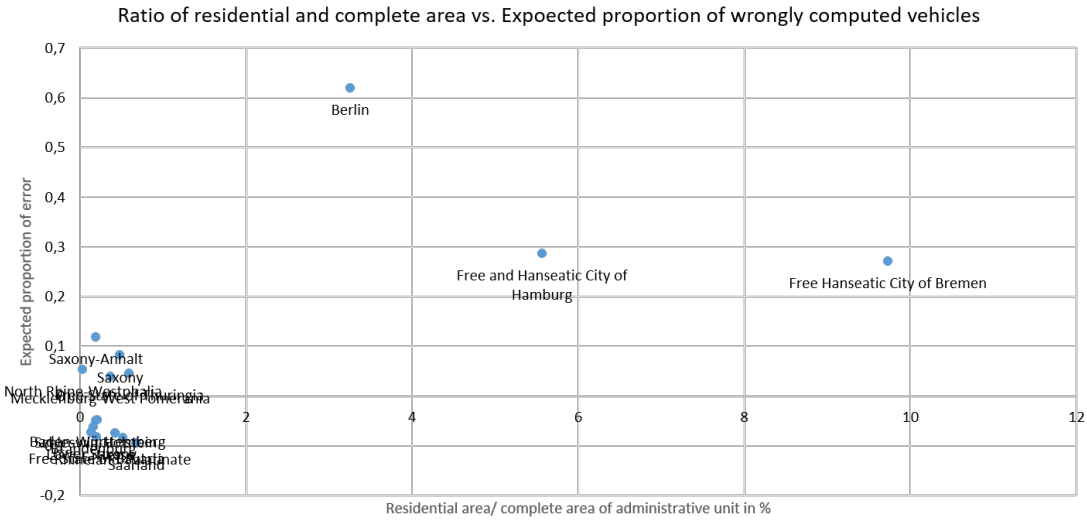


Figure 4.20.: Ratio of residential and complete area of administrative unit in percent against expected proportion of wrongly computed vehicles.

We restrict ourselves to the easiest assumption that the entries of p are represented as a linear function of the ratios $t_1, ..., t_{16}$ up to some small error:

$$p(t) = m \cdot t + b, \tag{4.6}$$

where m, b are chosen such that $p(t_j) \approx p_j^0$, $j = 1, ...16$. We use a linear regression model

$$p_j^0 = mt_j + b + \epsilon_j, \;\; j = 1, ..., 16, \tag{4.7}$$

and calculate the least squares estimates $\hat{m}$, $\hat{b}$ of m and b:

$$\begin{aligned} \hat{m} &= \frac{\sum_{j=1}^{16} \left(p_j^0 - \overline{p}^0\right)\left(t_j - \overline{t}\right)}{\sum_{j=1}^{16} \left(t_j - \overline{t}\right)^2}, \\ \hat{b} &= \overline{p}^0 - \hat{m}\overline{t} \end{aligned} \tag{4.8}$$

where

$$\begin{aligned} \overline{p}^0 &= \frac{1}{16}\sum_{j=1}^{16} p_j^0, \\ \overline{t} &= \frac{1}{16}\sum_{j=1}^{16} t_j. \end{aligned} \tag{4.9}$$

In the real data example of the German federal states, this results in

$$\hat{m} = -0,035016278 \tag{4.10}$$

$$\text{and } \hat{b} = 1,017340841. \tag{4.11}$$

The resulting values $\hat{p}_j = \hat{m}t_j + \hat{b}$ then lie between 0.68 for Bremen and 1.02 for North Rhine-Westphalia.
Note that we have (intentionally for simplicity) neglected the dependency caused by the requirement $x'_{1,1}+...+x'_{1,16} = 1$. The resulting $x'_{j,1} = \hat{p}_j x_{j,1}$ do not sum to 1, but only to 0.998. We correct this in a final step by setting

$$x''_{1,j} = \frac{x'_{1,j}}{\sum_{k=1}^{16} x'_{1,k}} = \frac{\hat{p}_j x_{1,j}}{\sum_{k=1}^{16} x_{1,k}}. \tag{4.12}$$

Alternatively, we could have used a constrained based least-squares estimate with the constraint

$$\sum_{j=1}^{16} x'_{1,j} = m \sum_{j=1}^{16} t_j x_{1,j} + b = 1. \tag{4.13}$$

As the unconstrained least-squares estimate already almost fulfills the constraint, this will not make much difference.

Figure 4.19 already included the expected differences in shares after the simulation with this new proportions as orange bars. Obviously, for most of the federal states the height of the bar is reduced. In some cases, like for Saxony-Anhalt or the already mentioned North Rhine-Westphalia, the error grows slightly but for the city states it is reduced significantly. For Bremen the error changes its sign, i.e. now few vehicles are expected to be simulated instead of too many, but its absolute value is lessened. For Berlin the new share is better but still far from optimal.
These declines might only result from the use of a linear approximation, but they show the general problem that improvements for some entities often are disadvantageous for other administrative units. Depending on the country considered, some different values for m and b might be more suitable also. Thus, the analysis has to be repeated multiple times. Even for the rural districts and urban communes a new calculation has to be performed. The overall correspondence between population and vehicle figures seems to be sufficient for the usage modeling, though the additional computational effort is expected to overshoot the enhancement. Section 5.1.3 includes an examination of the simulation results using the population only in detail. It is shown there that the obtained locations look quite reasonable. Thus we do not invest more time in the correction of the applied distribution here.

4.3. Determining distance distributions

After the simulation of the home locations of the vehicles the next required input of the usage modeling is the distribution of distances. These depend on various factors like the *region* in which the simulation is conducted as well as on the user

and trip purpose. Especially the driving patterns of commuters or taxis can be decomposed into differing types of single trips. The simulation of representative routes requires several settings for each of those. Only then the examined usage group can be mirrored adequately. In the description of the models given in chapters 2 and 3 two different kinds of bounds were used. First of all, applied for the light-duty commercial vehicles as well as for attendance and break activities of commuters for instance, upper limits were prescribed. In addition, for all other commuter and taxi trips, distance classes also lower bounds were used. The former can be included in the second concept by just using zero as missing value. Thus, in the following we mainly concentrate on the more complex type.
The distance distributions are required to create the circular rings introduced in sections 4.1.2 and 4.1.3 where target points are searched. They should be chosen in such a way that reasonable driven distances are obtained in the end. Thus, two inputs are requested. First of all, the actual distance distribution obtained usually from traffic surveys or other statistics. Secondly, a method to transform the usually reported driven distances into linear distances required for the simulation independent from the road network. Both issues are considered independently of each other. The first is rather easy to handle. In most countries there exists some ministry of transport regularly executing, or ordering from special companies, traffic surveys. Therein people are asked to record their trips during a predefined time period in combination with characteristics of the routes. These additional attributes usually include trip purpose, possible attendants, traveled distance and duration, means of transport and other pieces of relevant information. Summarized results are published in tabular form. There, the outcomes for combinations of various variables of interest are reported. Commonly, they contain distance distributions aggregated by trip types, general specifications of the origin or starting times. Thus, rough distance distributions can be extracted directly, more complex combinations have to be calculated by iterative proportional fitting for instance. This method is described in [23] or [45] for instance.
Examples for such traffic surveys include the *Enquête nationale transports et déplacements (ENTD)*[46] for France or the *Mobilität in Deutschland(MiD)*[38] for Germany. For the latter even some public use files exit including anonymized travel diaries. These allow more elaborate evaluations. A detailed description on this is presented in the next section. The tabular form can be applied nevertheless, often they are the only data available that is especially free of charge.
The translation of the distances classes included in the summary statistics of the traffic diaries can be performed based on detour factors already mentioned earlier. Rough values for them can be found in literature for selected countries, but we are also able to estimate them using VMC®. This topic is handled in detail in chapter 5.2.1. We also mention there how correctly determined detour factors could solve the problem of unavailable data.

4.3.1. Evaluation of MiD2008 data as an example of travel surveys

The main goal of the evaluation of traffic surveys is to gain the distribution of trip lengths. These essentially depend on the purpose of the trips and have to be determined correctly for the specific usage model of interest. Like already discussed, they are contingent on the starting point of the trip, on the available time or on the overall purpose of the complete journey. If people run on errands on their way to or back from work for instance, they often prefer stores in between their home and work place. If they go shopping on a day off, they are expected to accept a larger distance. Other people like doing their daily shopping during lunch break and are limited to a certain time span. These examples give an impression on the amount of distributions of trip lengths that have to be identified. Additionally, the frequency of the different trips has to be approximated reasonably.
In the following sections poll data collected in the framework of "Mobilität in Deutschland 2008" ([38]) is considered. This survey was conducted by the Institute of Transport Research of the DLR([2]) in cooperation with the social research institute "infas"([4]) by order of the Federal Ministry of Transport and Digital Infrastructure([1]). In this project people in Germany were, amongst other things, asked to report their trips on a single day.
Prepared data provided by the DLR Clearing House Report shall be used to determine several settings for the usage simulation of commuters. We roughly describe the different parts of the data set and sketch how relevant information can be extracted but we do not go too far into details.

Description of data

Information collected in the survey is summarized in the five main tables "Auto", "Haushalte", "Personen", "Reisen" and "Wege". These public usage files (PUF) were accompanied with one additional file each where data was cleaned but not checked for plausibility. In two sequential passes people were asked about attributes of their households like number and age of members or count and type of available means of transportation first. Afterwards household members were interviewed individually. In addition to personal characteristics such as personal life or health restrictions they were inquired of their trips on an appointed date and of journeys including at least one accommodation during the last three months.[37]
In order to guarantee anonymity, people can only be identified by a combination of household id and a number for each member of it. Their places of residence can only be reproduced by the name of the federal states they belong to and the settlement structural municipality type defined by the Federal Office for Building and Regional Planning (BBR) at different levels of aggregation. These include for instance information on the degree of urbanization of the considered region.
As a basic sample for the survey 25,000 households were selected reflecting the

German residential distribution in the federal states. Inside those, administrative units were classified by districts and settlement structure and chosen depending on their number of inhabitants. In order to obtain sufficient data for each category and level of detail, further survey participants were added. In combination with regional adjustments a total of 661 household belonging to 490 different communities was selected.[36]

Preselection of attributes

The data relevant for commuter simulation can be taken from table "Wege" conjoined with its supplementary file. In combination, these two data sets allocate about 134 characteristic values for 193,290 single trips. Restricting to daily or weekly tours, the remaining MiD tables are skipped due to their minor importance for the time being. In the case that further information on which brand or type of car is used for which kind of journey, they can be joined by common key variables without difficulty.
Our goal is to detect commonly carried out trips essentially classified by purpose, origin and destination. Afterwards, we want to analyze their proportion, sequence and particularly the distribution of corresponding driven distances. Additionally, the dependence on the home location of test persons shall be investigated. We want to restrict ourselves to car usage, thus not all reported trips are relevant. For that purpose most of the travel data is just used to determine journeys of interest whereas the actual evaluation is only performed only on a minor count of characteristic values.
First, we concentrate on 46 variables from the PUF and eight from the additional file. Some of them are combined such that the overall number of attributes per single trip is reduced to 34. In that process, the following rules were applied: Initially, the two identification numbers "hhid" and "pid" were combined to a single attribute. These serial numbers for households in the survey and persons belonging to those were liaised such that trips of persons can be filtered easier by just considering a single variable. Though, it is still possible to determine people living together. This new code number also enables the comfortable join of data from PUF and additional file.
Next, various logical variables indicating the usage of single means of transport like bikes, cars or trains on each trip were added and saved as a new parameter. Since we want to investigate the trips performed by passenger cars, the additional information that parts of the routes are conducted as motorist, is still kept. These two characteristics facilitate the exclusion of irrelevant trips. Subsequently, the possibility of attendance is examined. In the survey people were asked if they were on the move together. If they were accompanied by at least one child or adolescent living in the same household, this was registered and allows for example the identification of trips performed by parents taking their children to kindergarten or school before heading to their place of work.

Determination of trip types

In the MiD data different trip purposes are already included by classification of the target points. However, the usage modeling requires the combination with the origin. Only then the distance distributions needed in the simulation can be determined reliably. A way with purpose "shopping" for instance is further specified as daily or weekly commodity, which is of course also of interest in a detailed usage model where destinations are chosen from a diversity of POIs, but this is not the most relevant information. We rather want to know which trips in the simulation can be matched. Thus, the trip chain for each individual of the survey is passed, including all means of transport, and all origin-destination pairs are assigned. Hence, the ways from work to home can for instance be distinguished from those between leisure time activities and the home location. It is not possible to conduct this split directly on the original data. Additionally, the number of stops on the ways home can be estimated now, even separately for different types of activities. Thus, detailed distance distributions and corresponding frequencies of trip types can be obtained.
Routes with purpose attendance or breaks can also not be extracted directly from the traffic diaries. As already stated, the search for candidates for the former is rather easy. First, people are checked to be accompanied, afterwards the destinations for all individuals driving together are compared to find out who is escorting whom. Two possibilities are for example that children are brought to school or kindergarten, which is the trip purpose we are looking for, but alternatively adults are accompanying their children to shopping or other activities.
Breaks are even harder to identify. Here, also some time component is respected. One necessary criterion is that a work location is approached twice a day. However, no geographic coordinates are given, neither some hint if it is the same workplace in both cases. Hence, the time between leaving and returning is employed. Based on "common" working hours limits of 30 or 60 minutes are assumed to be suitable. In this process it is ignored when the break is taken. Thus, multiple shifts can be modeled and even those lasting from one day to the following are included. The identification of the activities during breaks obviously is the most problematic. The reliability of the resulting distance distribution is also put into question since only few entities are obtained.

Determination of distance distributions

The final definition of distance classes is then rather easy. Trip purposes or rather the origin-destination pairs are aggregated to groups first. Afterwards, for each category the reported driven distances are split into intervals. Here, an upper limit even for the last interval is required. Otherwise the computation of a circular ring during the simulation of target points is not possible. If some outliers are present, it is expected that these can be dropped without destroying the accuracy

of the distribution estimation. Additionally, also trip purposes can be combined if single groups contain too few data points. This is especially the case if the detailed classification of target points is used. It makes sense to put all shopping trips in one category since otherwise also in the simulation of the usage model a fine differentiation of POIs is required. On one hand this causes additional computation time in the preparation of the data. On the other hand the density of suitable target points is reduced. Then it happens more often that no adequate POI is located in the selected distance class, and the simulation of this vehicle has to be restarted. In the overall sum this could enlarge the computation time significantly.

5. Reliability of algorithm

Up to now we have considered the structure and components of different usage models as well as different data sources the simulation can be based on. However, for some of them the integration is not self-explanatory. Special methods are required. The most important are described and reviewed in this chapter. Additionally, we examine the influence of abort criteria restarting the simulation.

5.1. Simulation of home locations based on available population data

In this section we illustrate the simulation of home locations for commuters based on statistics in more details. We only consider population statistics here, but the basic procedure stays the same also if the vehicle distribution or some other data is available which is more suited for other types of usage models. In chapter 4.1.3 we already showed how specific coordinates in land-use areas can be simulated. In that process we distinguished the two cases that an initial point or a destination lying in some given distance class are searched. There, we assumed that the relevant regions have already been preselected. These result from the procedure described here. Recapitulating figure 3.2, we consider the central node "Choose unit". The determination of a commune is neglected since the outcomes can only be compared reliably on the level of federal states and rural districts in combination with urban municipalities due to lack of more details in the data. In the following, we assume that the corresponding level l of the administrative units has already been chosen.
The input of the simulation of the home location then consists of three major components:

- A list $AU = (AU_i)_{i=1:k_l}$ of all k_l administrative units of level l, given by their boundaries represented as *multipolygons* (compare chapter 4.1.6).
- The considered statistics of interest stored as discrete distribution summarized in $p = (p_i)_{i=1:k_l}$. p_i contains the ratio of the overall population that lies in AU_i. If the statistics include absolute numbers, these are converted such that $\sum_{i=1}^{k_l} p_i = 1$.

- A catalog of residential areas $R_{i,j}$ combining their geometric description with an indicator for the administrative unit they belong to. Thus, i still runs from one to k_l. The upper limit of j is not constant but gives the number of single residential neighborhoods in unit i. For each neighborhood $R_{i,j}$ also its area $a_{i,j}$ is already computed.

If required, also geometric centers $(c_i)_{i=1:k_l}$ of the administrative units or those of the single residential zones, $(c_{i,j})_{i=1:k_l,j=1:\#\text{residential areas in } AU_i}$, can be computed.
The requested result of this part of the simulation are the indices of selected administrative units. If for instance n home locations are required, a list $I = (i_1, ...i_n)$ is computed. Afterwards, all corresponding residential areas $R_{i_r,j}$ are extracted one after another for each $r = 1 : n$ and used as input for the methods given in section 4.1.3 in order to compute the n geographic coordinate pairs. Since we usually simulate the drivers one after another, we assume $n = 1$ in the following and drop the second index in the notation.

5.1.1. Methods to model population figures in other sources

The modeling and simulation of population distributions has been an active field of research for several decades. It is still of current interest since the ongoing development of geographical software enables the cost-efficient application of new methods based on by now available geographical data. We want to give an on no account complete overview of methods presented in literature that can be applied to create representative synthetic populations like we require them in the usage modeling.
Some of them, like Birkin and Clarke [23] and Bhat et al. [20], concentrate on microsimulations that produce large numbers of households and the individuals belonging to them according to joint distributions for several attributes like socioeconomic ones aggregated from different surveys. Unfortunately, they stop at the level of districts used in the always included census. Thus, even if the population is reflected well, these methods are not suited for our purposes. Lovelace et al. [45] start with the same approach but include some considerations about the determination of specific home locations inside the administrative units. They prefer the inclusion of population densities inside these units to obtain a more realistic distribution of locations than resulting from the simple technique of just placing all households at the zone centroids. Additionally they propose the application of shortest path algorithms for the determination of work places instead of the use of employment centers. However, they do not report methods for implementation. Beckman et al.[19] include public use micro-data samples containing true home locations. Such data is not available in our case.
There are also a lot of researchers that explicitly consider the construction of population densities based on geographical data. For that matter, multiple different methods to construct reliable and realistic distributions are proposed, but

the opinions about suitability, correctness and practicability differ strongly. The following excerpt of considered techniques again makes no claim to be complete, neither on papers dealing with them.

- **Pycnophylactic interpolation**: This method introduced by Tobler [58] at the end of the seventies seems to be the benchmark to rate most other techniques. The goal in this procedure is the creation of a smooth map based on packaged data [58]. A lattice is laid over the complete region such that in each administrative polygon at least one lattice point is placed. Then the whole weight of the population count is distributed over all points inside and shifted iteratively, such that smooth crossovers to neighboring units are created. In that process, the pycnophylactic property, this means mass-preserving inside the units, is guaranteed. Though, after aggregating the counts on the original level, the initial values are obtained. Additionally, non-negative population counts can be prescribed. Tobler also mentioned that individuals could be assigned to particular lattice points using Monte Carlo simulation. One critical point in this approach lies in the determination of boundary values. Counts assigned outside of a domain will effect the smoothness near the edges, the effect even propagates inside.[58] Thus they have to be chosen carefully. Some details on this approach can also be found in [34] and [61]. Additionally, Rase [49] enhances Tobler's pycnophylactic interpolation by using irregular triangular networks (TINs). He argues that this data structure is better suited to the polygonal shape of the administrative units than a simple equidistant mesh. However, this approach is not able to solve the problem of suitable boundary conditions. Especially Hay et al. [34] also note that it is unrealistic to assume no sharp boundaries in human population distributions, even if the method gives an elegant way to map a discontinuous to a continuous surface.

- **Areal weighting**: This simple approach is also used in comparisons often. Like in pycnophylactic interpolation, a regular grid is overlaid over the administrative unit. Then a population count according to the proportion of the polygon area inside the raster grid is assigned. It has the advantage of being rather simple, but human population is not to distributed uniformly in space.[34] Details on this method can also be found in [33].

- **Dasymetric mapping**: "A dasymetric map depicts quantitative areal data using boundaries that divide mapped area into zones of relative homogeneity with purpose of best portraying the underlying statistical surface".[31] It "uses ancillary information [...]at higher spatial resolution than the population polygon data to help allocate population [...] who are assumed to differentially inhabit land-use types".[34] The zones have boundaries resulting in sharp changes in the mapped statistical surface and thus are halfway between a smooth and stepped representation.[31] This approach is also rather simple and requires little additional information, but has the disadvantage

that weights to possibly different land-use classes have to be defined.[34] According to Gregory [33], it suffers from less error than areal weighting when data between source and target areal units has to be transfered.

- **Smart interpolation**: This method expands the dasymetric mapping in the way that it incorporates a huge variety of ancillary data. It is based on the fact that humans distribute themselves non-randomly in an environment but rather live near roads, not on top of mountains, etc. It tries to respect the geography of the area explicitly. Thus it can be highly complex and needs good estimates on the required weights.[34]
- **Area-to-point kriging and other geostatistical methods**: The summary of these methods all reproduce areal data when re-aggregating the computed surface to the original units. Physical boundaries can be prescribed as well as inequality constraints like the non-negativity of population counts. The geostatistical methods have the advantage that the uncertainty in each point value of the created surface can be assessed. Additionally, Tobler's pycnophylactic interpolation is included as a special case as well as common methods like choropleth maps or kernel smoothing.[61]
- **Choropleth mapping** and **kernel smoothing**: These two traditional methods shall only be mentioned for the sake of completeness. In the former the inhabitants of an administrative unit are considered as volumes and prisms are visualized over their polygonal boundaries. The heights of these prisms are computed as the volume divided by the polygonal area.[49] In kernel smoothing, a smooth surface, free of abrupt chances at boundaries and without the assumption of homogeneity inside an area is created. However, here the support differences between the source data and the prediction surface are not properly taken into account.[61]

As already mentioned, in the cited papers some of the methods are compared to each other but no common favorite can be found. Yoo et al. [61] for instance prefer geostatistical methods before pycynophylactic interpolation since the obtained results are comparable but the statistical techniques can be applied in stochastic simulation. Hay et al. [34] prefer areal weighting as easy to implement method that is most accurate. It is stated that pycnophylactic interpolation always decreases accuracy and is only justified by aesthetic reasons. Smart interpolation is also not preferred since it strongly depends on the spatial resolution of the employed GIS data. Dasymetric mapping is considered as problematic there, due to the definition of relative weights. Gregory [33] instead favors dasymetric mapping as producing less error compared to areal weighting.
Since land-use information can be easily extracted from the VMC® database, we expect some kind of dasymetric mapping to be a good choice. Eicher et al. [31] compare three slightly different alternatives for this. The first two are conducted both with polygon and grid form of land-use data. Their difference cannot be verified statistically but, based on the examples used, the polygon versions are

usually better. Since we get this type of data directly from the database, we see no advantage is converting it to a grid form, either. The three methods can be distinguished in the following way according to [31]:

- **Binary method**: In this technique the population data of each county is assigned to only urban and agricultural or woodland cells. Thus a binary labeling as inhabited and uninhabited zone is conducted. This approach is the simplest of the three.
- **Three-class method**: Here, some weighting scheme is applied to distribute the population counts over three different land-use types within each county. In addition to urban and agricultural or woodland zones, the third category of forested areas is included. The sum of all zones of each type are filled with a fixed proportion of the population count for the county. In the cited paper, shares of 70%, 20% and 10% are used. Again, this method is easy to implement, but the proportions are hard to determine correctly. The given values are just estimated and actually have to be adapted properly. Depending on the distribution of types, small zones might get too much importance.
- **Limiting variable method**: In this most complex technique, as the name indicates, some limits on the population densities, in the sense of number of inhabitants per km^2 in the area, are prescribed. It also works with the three land-use classes and starts with single area weighting in the first step. Afterwards, some upper limits for the three types are fixed and the zones violating these limits are adapted. The spillover is evenly distributed to the remaining zones of the county.

A comparison of the three methods in [31] yielded that the last method was the one producing the least error before the second and the first technique. However, the binary method is the only one not requiring additional input. The others need some kind of prior computations determining the proportions and limits optimally. Since we have no indication for good values, we want to avoid these expensive calculations. Kim [40] introduces a hybrid technique combining the binary method with pycnophylactic interpolation which requires even more computational effort. Additionally, we decided earlier to concentrate on residential neighborhoods only for the simulation of commuters' home locations. We ignore retail areas including possibly also flats and assume that industrial and commercial zones are uninhabited in the sense of not containing residential properties. Hence, the binary method seems to be the most suited and is taken as a basis. The algorithm explained in the next section does not contain the reconstruction of the population distribution explicitly. It directly deals with the simulation of home locations which is required in the usage model simulation.
There, also a hybrid technique merging binary method and pycnophylactic interpolation is introduced. For our purposes however, it is also not optimal and thus neglected.

5.1.2. Description of algorithm applied in usage modeling

The modeling of the population distribution should fulfill several demands. First of all, it has to be efficient in the way that it does not require too much computation time to be constructed. Alternatively it could be stored in the VMC® database after it was created once, but then loading it should be fast. Even if this is possible, the construction has to be redone for every new *region* in the usage model, compare chapter 2.2, and thus is usually conducted several times. Secondly, the algorithms should be flexible such that they can directly be applied to the new *region* without further preliminary work. The last and most critical point is the random selection of locations from the model. This step has to be performed for every run of the usage model simulation. Hence, the computation time of this step is multiplied by factors in the scale of several thousands for each population creation. A large amount here could thus make the whole procedure inapplicable. These specifications also support the use of some kind of dasymetric mapping. The actual algorithm employed to simulate a single location conducts the following steps:

1. All administrative units AU_i and their corresponding relative probabilities p_i for the chosen administrative level are extracted from the database.
2. Using the discrete version of the inverse transform method, the index i_{chosen} of a single unit is selected randomly. In detail, a random number u is generated for the uniform distribution $U(0,1)$. Then the index i_{chosen} is determined as the single index fulfilling $\sum_{j=1}^{i_{chosen}} \leq u < \sum_{j=1}^{i_{chosen}+1}$.
3. All residential areas $R_{i_{chosen},j}$ contained in the selected administrative unit as well as their surface areas $a_{i_{chosen},j}$ are extracted. The new overall sum $a_{i_{chosen}} = \sum_k a_{i_{chosen},k}$ is used to determine the new proportions

$$p_{i_{chosen},j} = \frac{a_{i_{chosen},j}}{\sum_k a_{i_{chosen},k}} = \frac{a_{i_{chosen},j}}{a_{i_{chosen}}}$$

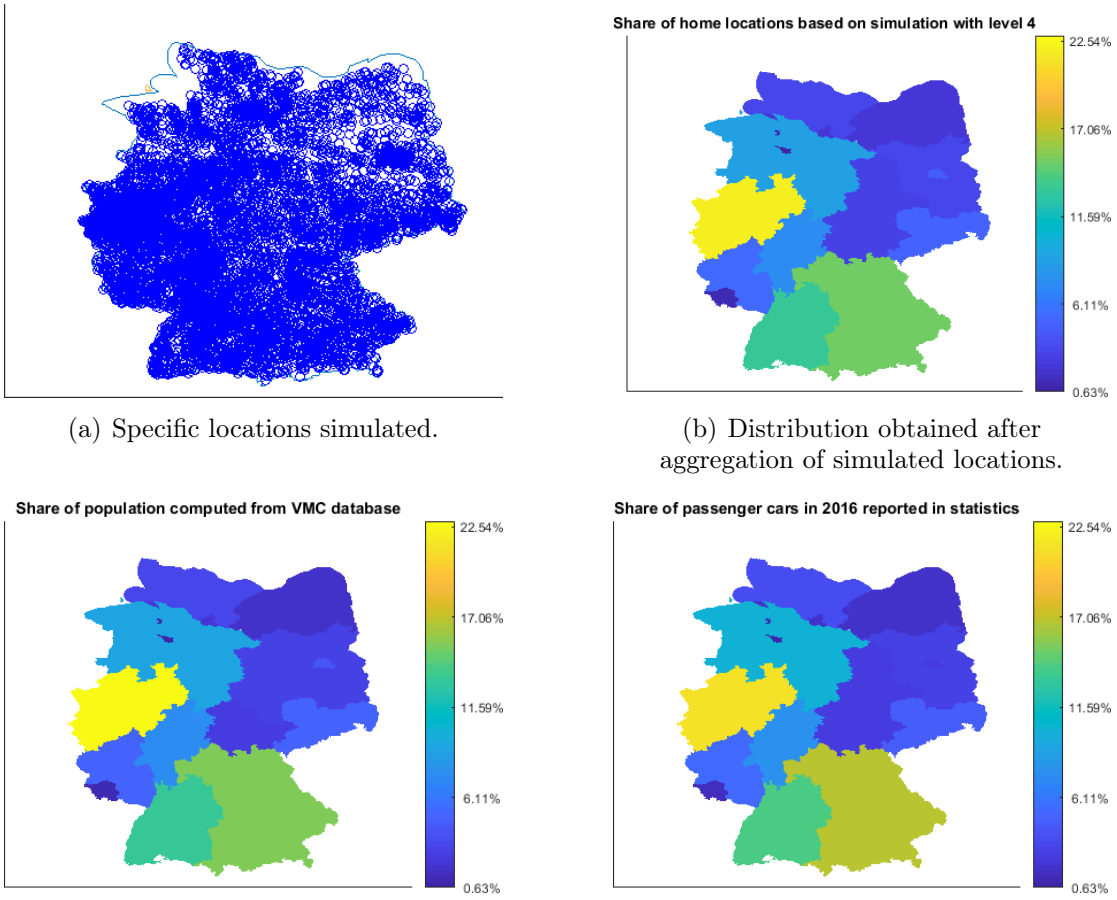
   of the single zones.
4. The discrete inverse transform method is repeated with this proportions and a neighborhood $R_{i_{chosen},j_{chosen}}$ is selected.
5. Inside $R_{i_{chosen},j_{chosen}}$, a home location is simulated. Here, the methods for ROIs presented in chapter 4.1.3 are applied.

The last two steps already have been described in the mentioned chapter 4.1.3 but are repeated to illustrate where some kind of dasymetric mapping is performed. Obviously, this algorithm fulfills the pycnophylactic constraint due to the first selection procedure. Of course, a pure dasymetric mapping would consider all $R_{i,j}$ directly without choosing AU_i first. Though, then a count is assigned to each zone $R_{i,j}$ and summing them up gives the population count for the whole

region. The application of a random number generator then implicitly computes a frequency distribution by dividing the single counts by their sum resulting in the same $p_{i,j}$ that are obtained in our approach. We prefer this two-step method since there exist lots of residential areas, thus the unit interval is split in many parts. For Germany for instance more than 267,000 neighborhoods exists. This could promote rounding errors and blow up the computation time. The number of p_is is comparable small and the amount of $R_{i_{chosen},j}$ can be handled better then the complete data set.

5.1.3. Comparison of simulated and true distribution

The verification of the employed algorithm shall be done graphically. Therefor multiple simulations on different administrative levels and with varying sample size shall be conducted. We start with the comparison for a simulation of 10,000 home locations using population statistics on the federal state level as input of the algorithm. Figure 5.1 first depicts the specific simulated locations. These are assigned to the federal states they lie in and the resulting proportions are shown in the second graphic. In order to compare the distributions, also the population computed from OSM as well as the official vehicle statistics are presented with the same colorbar. Obviously, the proportions used in the algorithm are treated correctly. The distribution obtained after aggregation can hardly be optically distinguished from the population figure. Additionally, the difference in densities can already be detected from the marked locations. Especially in eastern Germany the city of Berlin can be identified inside a large area containing clearly less simulated homes.



(a) Specific locations simulated.

(b) Distribution obtained after aggregation of simulated locations.

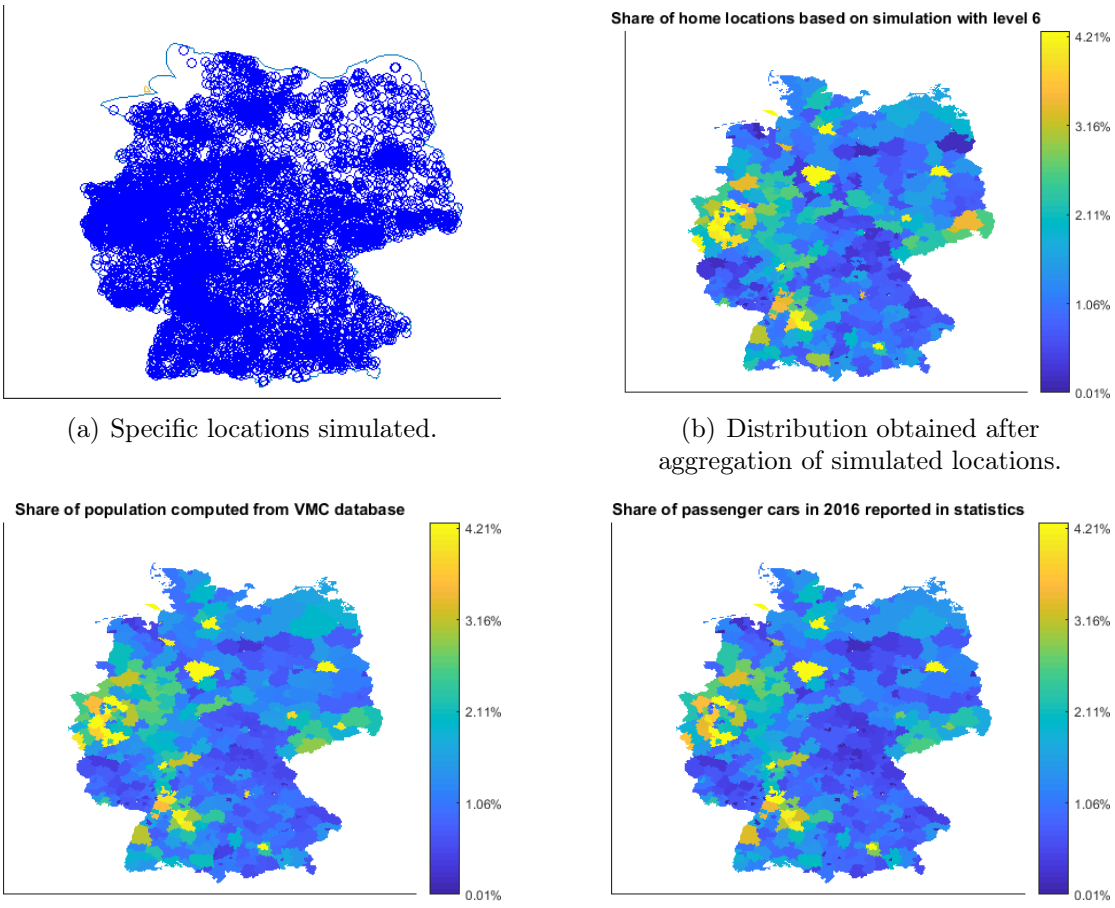
(c) Population distribution computed from VMC® used as input.

(d) Vehicle distribution given in statistics.

Figure 5.1.: Comparison of distribution obtained from 10,000 simulated home locations and input statistics based on federal states.

The same results are obtained for the evaluation for the simulation based on the smaller units of rural districts and urban municipalities shown in figure 5.2. Again, the overall distribution is reflected well but single units are defective, especially in eastern Germany, where one region got a too small share and one was rated too high, but also in West-Germany where there is one area which was nearly ignored. However, this results from the random number generator used in the algorithm. Now we know that the algorithm works well for the two different administrative levels. The level of federal states respectively of urban districts and municipalities. What is still open for discussion is the preference for one of them if the only request is the usage modeling in the complete country. Does it make a difference which level is used concerning accuracy and computation time? An easy answer on this question is not possible. It can be said that there was no apparent difference in the time needed to chose the 10,000 home locations comparing multiple runs. The number of residential areas handled in the second selection step is ignorable. In the

OSM database there exist neighborhoods that range over multiple administrative units and have to be split, but the overall count of zones to consider grows only about one percent from federal states to rural district and urban municipalities. In the first random choice, 16 units are enlarged to 402 units when considering the smaller entities. This also has no serious effect on the computation time.



(a) Specific locations simulated.

(b) Distribution obtained after aggregation of simulated locations.

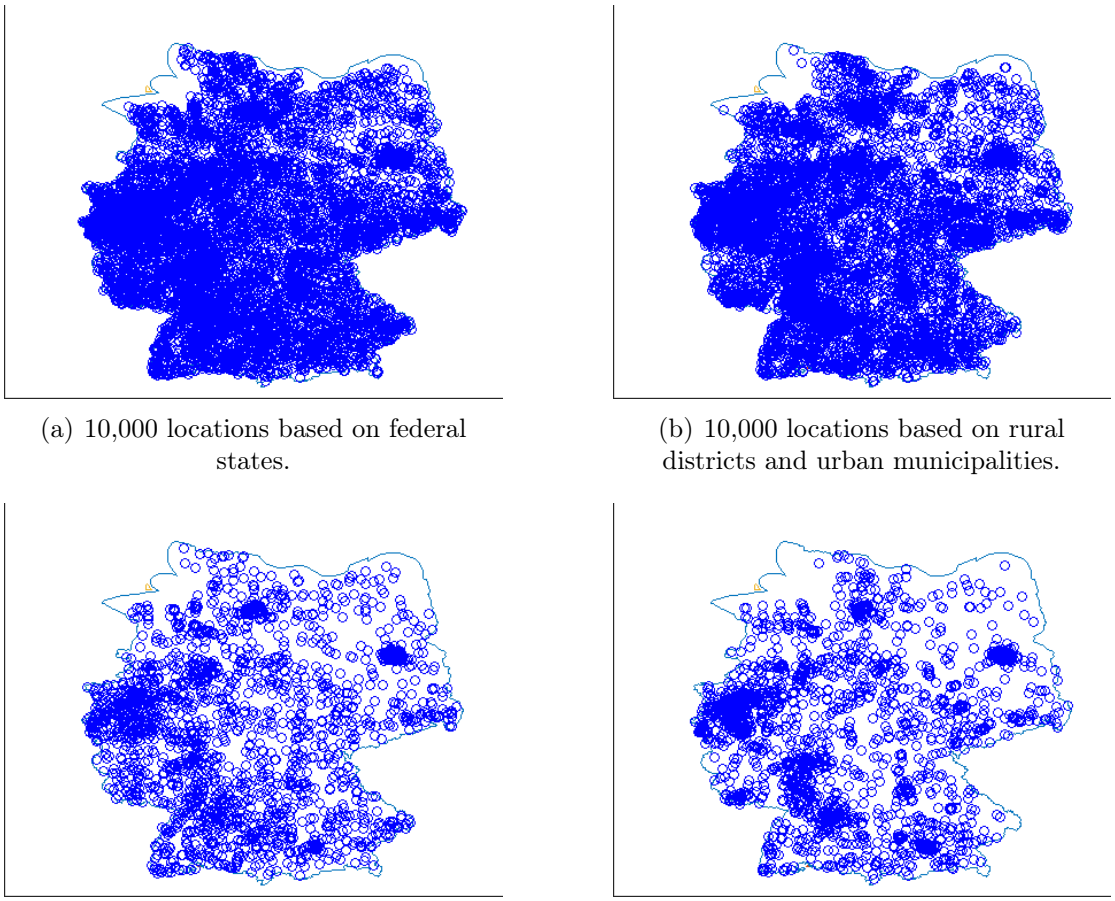
(c) Population distribution computed from VMC® used as input.

(d) Vehicle distribution given in statistics.

Figure 5.2.: Comparison of distribution obtained from 10,000 simulated home locations and input statistics based on rural districts and urban municipalities.

A significant difference between the two levels is found in the distribution of the home locations inside the units. Figure 5.3 depicts the results for simulations of 10,000 and 2,000 homes for both levels. The graphics based on federal states show a more uniform distribution of points over the country. The areas of high population density can still be identified, but the concentration is more smoothed out. Especially when the plot for the lower number of simulated locations are

compared, it can be seen that there are more "white spaces" that means more regions that are not included in the result. Which outcome is more suited depends on the purpose of the usage modeling. If the number of simulated vehicles is low, the federal states allow a better coverage of the country and thus potentially include more different roads traveled in the end. However, using the smaller units presumably reflects the vehicle distribution in more detail even if several districts are only included for a growing sample.



(a) 10,000 locations based on federal states.

(b) 10,000 locations based on rural districts and urban municipalities.

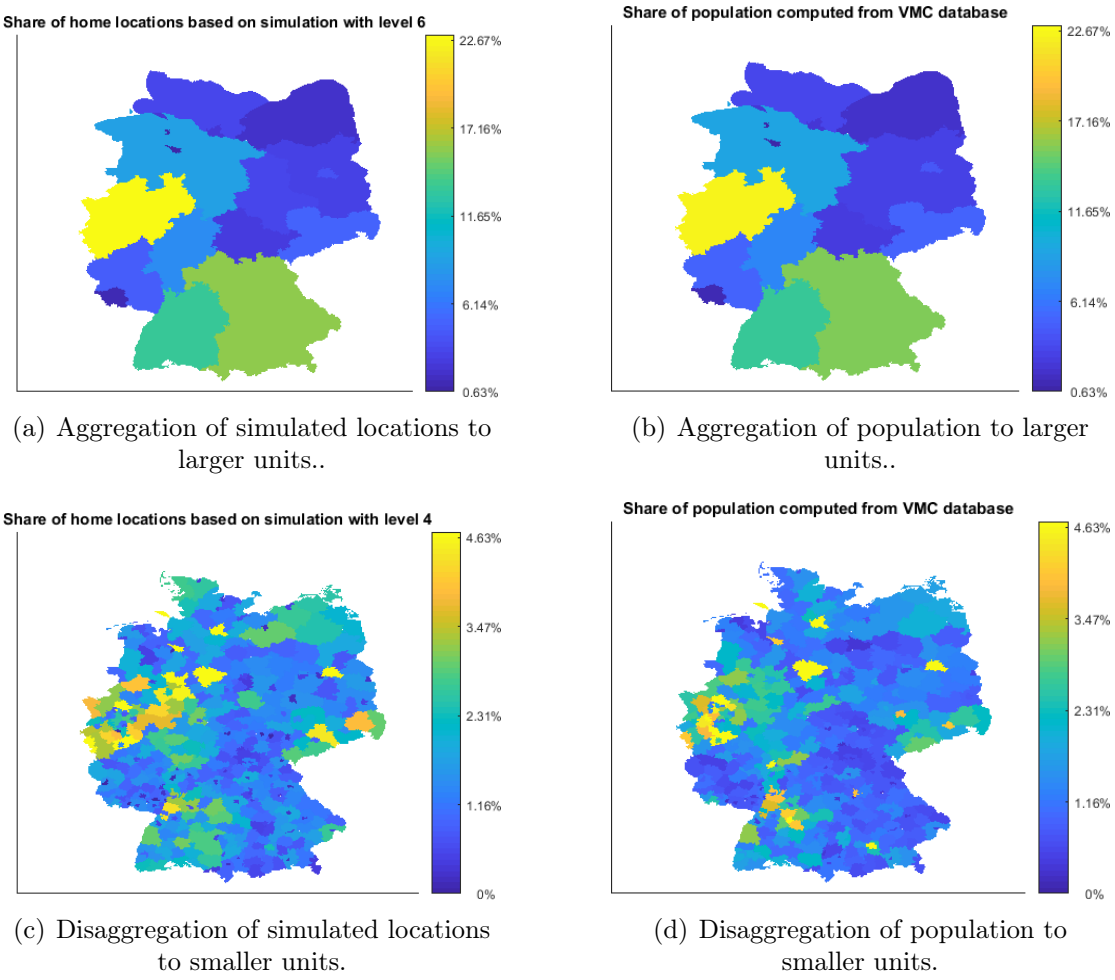
(c) 2,000 locations based on federal states.

(d) 2,000 locations based on rural districts and urban municipalities.

Figure 5.3.: Comparison of distribution of 10,000 and 2,000 simulated home locations. The left pictures used federal states as basis, the right ones used rural districts and urban municipalities in the first choice.

Additionally, the administrative level of detail for the final evaluation is important. Using rural districts and urban municipalities allows the aggregation on the federal state level. Though, disaggregating the larger administrative units produces an error in the population distribution, see figure 5.4. The accordance of the upper

graphics is quite good, the disaggregation yields the expected errors. Especially in North-Rhine Westphalia the smoothing of the density leads to higher proportions in more rural zones. The same holds for the area around Stuttgart where part of the population is moved away from the city.



(a) Aggregation of simulated locations to larger units..

(b) Aggregation of population to larger units..

(c) Disaggregation of simulated locations to smaller units.

(d) Disaggregation of population to smaller units.

Figure 5.4.: Comparison of results after aggregation and disaggregation.

Thus, the goal of the usage modeling and the number of simulated vehicles is essential. Due to the possibility of aggregation without remarkably growing computation time, the use of the smaller units is proposed. Though we know that the availability and preparation of data of this level might be a drawback, only allowing the use of the coarse division of the country into federal states.
The last point that shall be mentioned here is that the pure dasymetric mapping suffers from the same problems since there also a basis for the initial weighting has to be chosen.

5.2. Conversion between driven and linear distances

In the last chapter it was sketched that distance classes and their proportions are essential for the simulation of usage models. It was also mentioned that available data resulting from traffic diaries cannot be integrated directly but has to be preprocessed. That is the topic of this section. The applied factors are known with different names. In literature the notation "circuity factor" has been established, though the description as "detour factor" is more intuitive. Since we base our considerations on papers introducing the first notion, we use both names synonymously in the following.

5.2.1. Estimating circuity factors

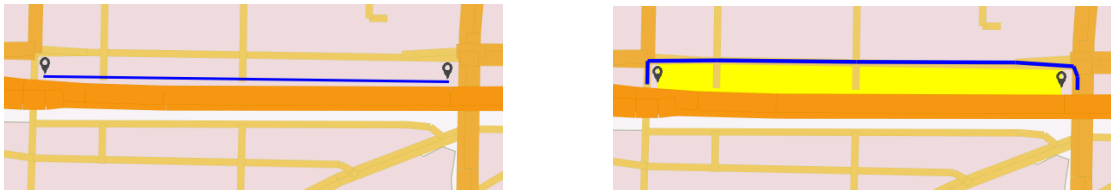
Available traffic surveys still require some further work before they can be included in the simulation of the usage models. People usually report driven distances whereas the algorithms for the simulation of target points previously described can only handle airline distances measured on the earth's surface. Thus, some translation between both measures is needed. Ballou et al. [18] introduce the concept of "circuity factors" (CFs) defined as the ratio between both corresponding values. They even provide some mean values and standard deviations for multiple countries of the world. In section 5.2.2 it is shown how such values can be used directly in order to compute fairly reasonable distance classes for the simulation. Obviously, the final results are improved using these values, instead of neglecting any translation, but the proposed factors are presumably too vague for larger countries. Since they are assumed to depend on the topology and traffic facilities, thinking of Germany for instance, rural, mountainous and urbanized regions are expected to have different factors. In [18] only a rather small number of sections of roads is considered and it is not shown how these parts are distributed over the whole country. Hence, the given values can be taken as a basis but not as optimal. Using VMC®, route dependent circuity factors can easily be computed and different influencing factors can be evaluated. As a start, the provided values are used to simulate a large amount of trips. Afterwards, these are routed, segmented and analyzed. Hence, a huge amount of road segments including characteristics like road type is obtained in combination with their geographic position. Thus, a more accurate analysis is possible. If regional or other significant differences are found, those can improve further simulations. Apart from this, the computation of factors using the same software and parameters as for the final evaluation generates the best and most consistent results. The complete procedure is explained in detail on the example of commuters in Germany in the following.
An additional benefit of using circuity factors is the possible workaround in case of missing data. Sometimes it is hard to find reliable distance distributions. Then,

often values are guessed or those of "similar" countries are applied. The use of circuity factors could also enhance this procedure. Here, a country with available mobility statistics is taken as a basis and the circuity factor is used to adapt the distance to the actually considered region. Hence, the topographical characteristics and the quality of the road network in the target area are taken into account. Depending on the availability of statistics and circuity factors the simulation of a usage model starts with at least one iteration of preconditioning. The results are analyzed regarding the specific circuity factor which is then applied in the true simulation afterwards.

Algorithm for determination of circuity factors

After the rather rough description how to compute better circuity factors, the procedure is explained in detail on the example of commuter simulation in Germany. As a basis, the mean value of 1.32 and standard deviation 0.95 given in [18] are taken. With them, the distance classes in the commuter poll are transformed like described in chapter 5.2.2. Afterwards, the simulation of the usage model is started with default values. Later on, influencing factors of the CFs shall be determined, therefore the number of commuters simulated should be large enough to provide reliable values for all types of trips. On the other hand, the computation time for the model simulation as well as the processing with VMC® cannot be neglected. As a good balance, the driving schedules of 4,020 commuters in Germany are created leading to routes on 20,100 days and an even higher number of single trips that can be analyzed in detail. Applying the standard settings in the commuter model, a number of more than 68,000 was obtained for the example data. Here, of course many tracks are included multiple times since people use to take the same way for regular trips without unusual incidents. These are only considered once in the evaluation, still keeping more than 36,000 different routes. The number of commuters is chosen in the way that in each of the 402 rural districts and urban municipalities in Germany exactly 10 vehicles are places. Of course, then the population distribution is violated, but in doing so the availability of a sufficient amount of trips in each region is guaranteed. Otherwise, there can be areas in Germany for which no reliable estimation of the detour factor is possible. Before the kml-files produced as outcome of the simulation of the usage model can be processed with VMC®, according settings have to be determined. It is recommended to employ the same factor model that should be applied for the true evaluation later on. The distinction of road types for instance influences the segmentation performed and might lead to slightly different trip lengths for the same route with different preferences. Surely, those imbalances are expected to be of small magnitude, but they might bias the obtained circuity factors. Since the settings have to be adapted at least after the preprocessing and before the true evaluation with VMC® is started, this step should be conducted in advance in order to get consistent values. Some details on the preferences that can be set

are illustrated in the concrete examples of chapter 6. Elaborative descriptions on factor models can be found in [52]. In short they form a concept to easily structure data by dividing it in different cells.
After the setting of all preferences, several steps are performed with VMC® for each driving schedule. First of all, the kml-file is loaded and the single trips represented by their starting and target points are routed. At the same time, the provided coordinates are mapped to the nearest feasible road. This is inevitable due to the characteristics of the input data of the simulation. POIs for instance are tagged at their actual places which are not necessarily directly at the adjacent road. ROIs might include streets but those are not considered in the random selection procedure since attached property boundaries of single estates are less meaningful than areas. The assignment by VMC® also has the advantage that multiple proximate candidates can be rated. At corner properties for instance, smaller streets can be preferred against busy ones. It is prevented that trips start on highways directly for instance, if a neighborhood is settled close to it. Figure 5.5 illustrates this circumstance on the example of a route parallel to motorway A100 near Bundesplatz in Berlin. Here, the points are mapped to the next smaller roads although their distance to the highway is shorter.
Afterwards, the shortest connection between consecutive points is computed. Again, this setting has to be consistent to the analysis of the actual usage simulation. The alternative of calculating the fastest route obviously leads to different distances traveled and thus different circuity factors.



(a) Original position of points

(b) Route after mapping to road network

Figure 5.5.: Mapping of points to adjacent smaller routes instead of highway A100 in Berlin close to Bundesplatz.

Identification of influencing factors

The result of this preconditioning iteration consists of thousands of single trips aggregated with various factors. The most important are the driven distance computed with VMC® and the straight line distance on the earth's surface. Their quotient gives the circuity factors for each trip. Additionally, for every route the initially simulated points prior to the mapping on the road network and corresponding expected linear distances are available. Figure 5.6 shows that both values are nearly everywhere the same. Only for small distances the differences

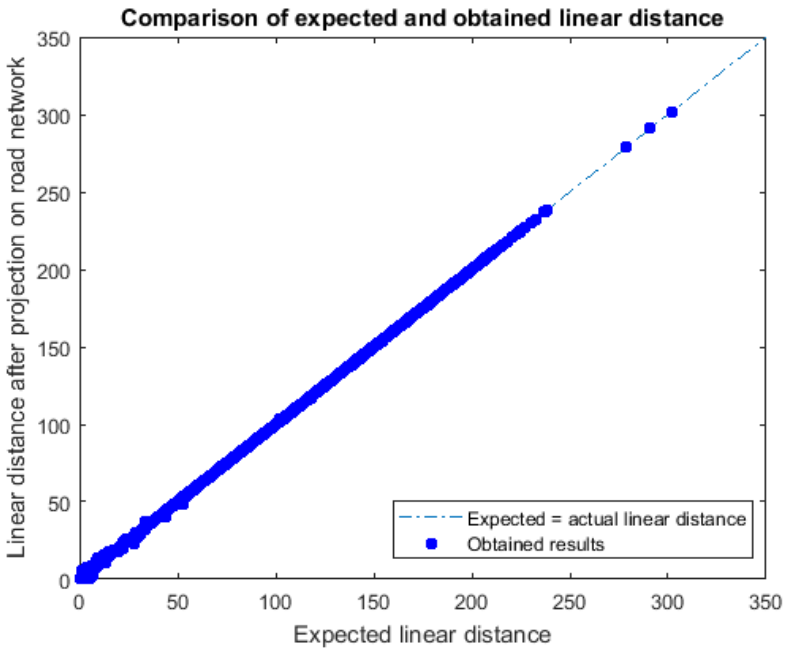
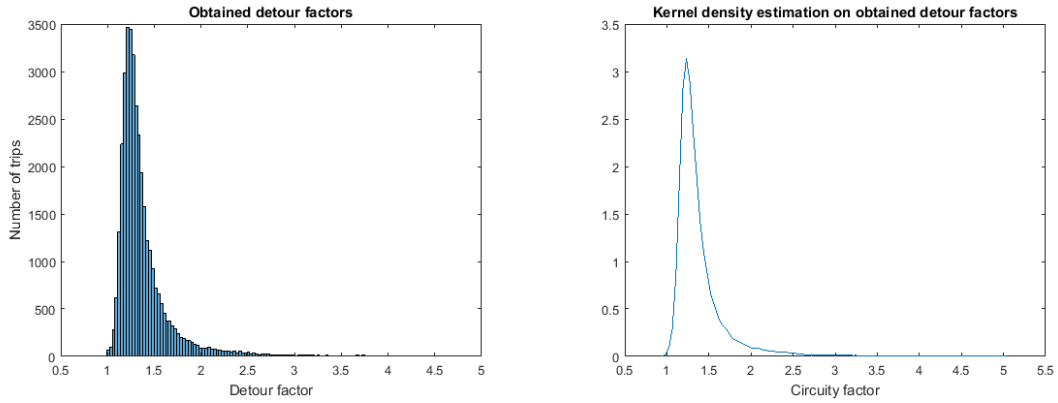


Figure 5.6.: Comparison of linear distances expected in simulation and obtained ones after projection of locations on road network.

look more severe. Calculating the relative errors with respect to the expected distances yields some outliers suffering from great disagreement. Though, the central 95% of the errors range from -2% to 2.7% . This is sufficient for our purposes. We prefer the use of the actual obtained distances in the following, if possible, because they base on identical coordinates like the driven distances.

Circuity factors are now calculated as the fraction $\frac{\text{Driven distance}}{\text{Linear distance}}$. They have a natural lower bound of 1 since the distance on the road network cannot be shorter than the straight line between the points. Due to some missing bridges in the map data and the avoidance of ferries in the routing, the simulations achieve an upper bound of about 36.1 which is really large. This limits strongly depends on the region considered and varies depending on the chosen points. Even in smaller areas of interest large detours are inevitable when the road network is poor or multiple one-way streets are existent. In our case this huge value forms an outlier that could be excluded from the next evaluations. The 99% quantile of the obtained detour factors has a value of 3.57 and even the 99.9% quantile is less than 10 and thus considerably smaller than the maximum. Hence, we now concentrate on those factors between one and five since we expect them to be reliable. Additionally, this reduction makes the next graphics more readable. Having a look at the histogram of the obtained CFs given in figure 5.7(a), it is suspected that the factors are not only influenced by effects similar for the complete region but also depend on some other parameter. The estimation using a kernel density estimator shown in figure 5.7(b) supports this assumption.



(a) Histogram of obtained CFs (b) Kernel density estimate for obtained CFs

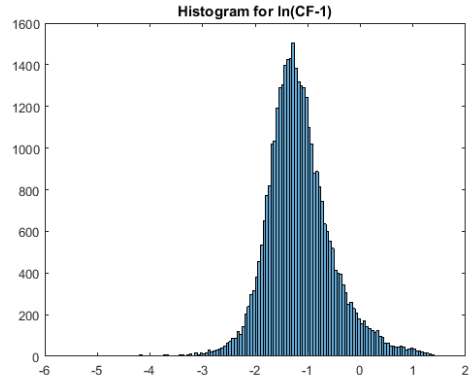
Figure 5.7.: Visualization of obtained circuity factors.

Obviously, the standard assumption of a normal distribution can be dropped directly. However, the strict lower natural limit of one makes a transformed log-normal distribution a promising candidate. In this process we try two types of transformation. First, we apply the natural logarithm directly, in the other case we first shift the detour factors by one. For the latter values $CF - 1$ the following relation holds:

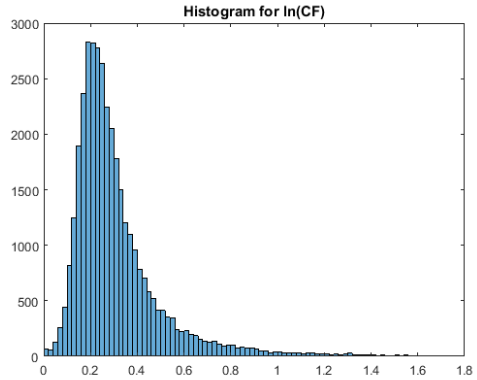
$$\begin{aligned} CF - 1 &= \frac{\text{Driven distance}}{\text{Linear distance}} - 1 \\ &= \frac{\text{Driven distance} - \text{Linear distance}}{\text{Linear distance}} \\ &= \frac{\text{Detour}}{\text{Linear distance}} \end{aligned} \tag{5.1}$$

Hence, it gives the additional amount of covered distance in relation to the linear distance. Here, we have to exclude all factors attaining the lower bound for which the logarithm is not defined. Since for these routes no detour is reported, this means we just leave out the rare events that can only occur if origin and destination are connected by a straight road. Due to the street network, this is expected to be achieved just for short trips.
For the two transformed samples we obtain the histograms given in figure 5.8. Both are still skewed. On a first glance, figure 5.8(b) would come from a log-normal distribution, whereas figure 5.8(a) resembles a mixture of two normal distributions, where the majority of data have mean≈ 0. Testing the hypothesis of a log-normal distribution with an Anderson-Darling test, in both cases the null hypothesis is rejected on the 5% level. We thus follow the visual hint that the logarithmic data might be a mixture of normal variables, and we try to divide the trips in groups of similar behavior in order to model the overall distribution of detour factors as a mixture of multiple components. We expect that the original data follow a log-normal distribution, the shifted and logarithmized ones a normal

distribution respectively. This assumption is based on the shape of the histogram just considered. It is supported by the quantile-quantile and normal probability plot given in figure 5.9. The middle 98% of the transformed values seem to follow a normal distribution. Though, the original values could result from a shifted log-normal distribution. Applying the Kolmogorov-Smirnov-Test to the data limited by these bounds, we are still not able to prove the assumption of an underlying single log-normal distribution. Accordingly, we have to group the obtained CFs by more complex criteria.

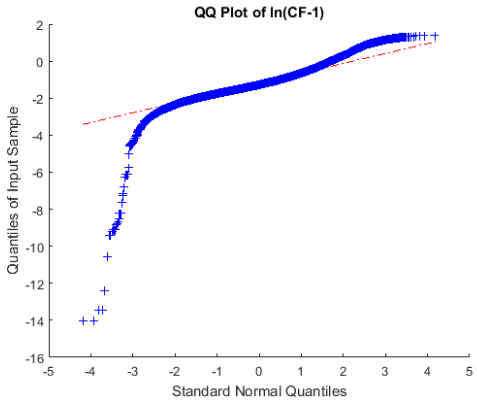


(a) Histogram of ln(CF-1), cut on the left to emphasize to obtained shape.

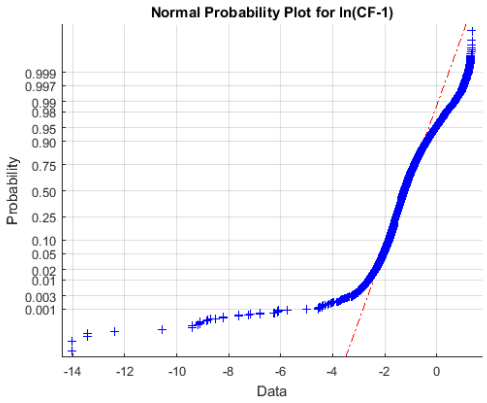


(b) Histogram of log(CF).

Figure 5.8.: Histograms for the two transformed samples.



(a) Quantile-quantile plot for ln(CF-1).



(b) Normal probability plot for ln(CF-1).

Figure 5.9.: Comparison of log(CF-1) with normal distribution based on quantile-quantile and normal probability plot.

In order to determine the main influencing factors we have more than 13 attributes with multiple different values available for each route which may be used

for splitting up the sample. Since the amount of data is huge, statistical tests for differences shall not be performed for all of them. We rather start with some visual check reducing the candidates. Therefore we again plot the linear against the driven distance and employ some group-coloring. Figure 5.10 shows the scatter plot obtained for a split by federal state. The dotted line marks the lower bound of a detour factor of one. The dashed line has a slope of the mean detour factor computed from the sample routes. For small distances, it is rather longer than the prediction-based on the dashed line, i.e. using the overall mean of the individual circuity factors. It reflects the relation between driven and linear distance reasonably well in the intermediate range, and it overestimates driven distance considerably for distances beyond roughly 100km. Note, that the points in the scatter plot lying considerably above the desired line essentially come from only some of the federal states.

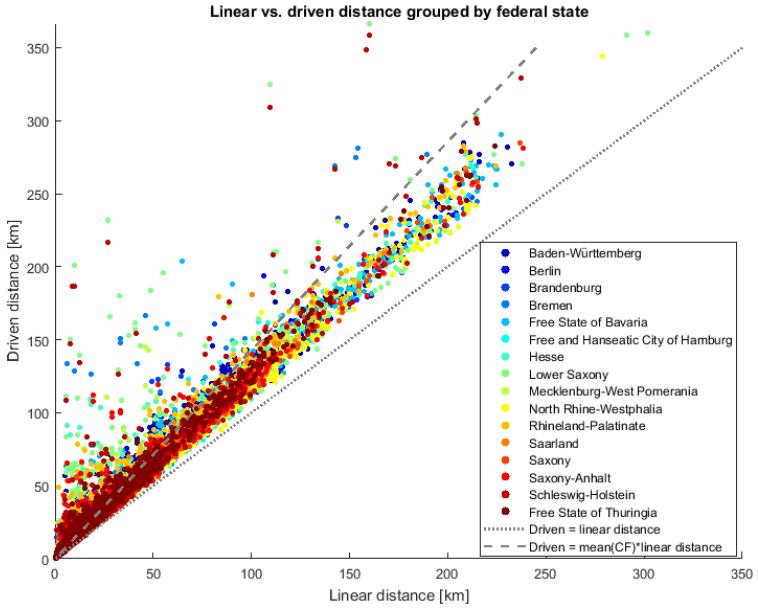


Figure 5.10.: Scatter plot of linear against driven distance including coloring of data points based on federal states.

These observations are not surprising at all. Without natural obstacles like a river or mountains, it is expected that for long distances, the ratio between driven and linear distance is closer to 1 than for short distances as the driver will have more options for planning his route to be as short as possible. For driving to a nearby location, on the other hand, often only one route is available anyhow which may include a considerable detour. Of course, this phenomenon also depends on the density of the road network which may vary between states. As a first attempt, we therefore split the sample into different parts corresponding to distinct intervals of linear distance, and we attribute an own circuity factor to each subsample.

We work with the shifted and logarithmized circuity factors only, because those are expected to follow a normal distribution if the groups are chosen correctly. A combination of analysis of variance and multiple comparison test with a significance level of 5% were performed. The required groups are obtained by splitting the driven distances of the routes into several intervals. The separation in three, four and even more groups showed that shorter and larger distances can be partitioned on different grids, but the medium part should be assigned to one group. We tested the upper bounds of 20km, 35km and 50km for instance and examined that the difference between intervals [20,35) and [35,50) cannot be shown. Two-sample t-tests performed on different splits of the data into two groups yielded that the interval limit should lie between 20km and 100km. Otherwise the null-hypothesis could not be rejected. Thus, there are various possibilities how to define new classes that influence the detour factors. The big advantage of this investigation is that the required information for the assignment to a group is available in the simulation.
Since no reliable effects resulting from other attributes of the routes can be found, the new knowledge on the circuity factors can be exploited easily on the simulation of usage models. We recommend that the number of separated groups should be based on the applied statistics. The outcomes of traffic surveys usually are summarized for distance classes. These can be used as initial configuration of the split. The routes created in the preconditioning iteration are then assigned to these classes. Again, statistical test have to be performed then. If the sample sizes in some intervals are too small, the groups can be combined. Depending on the results of the test, a number of mean values and standard deviations is computed then. Their count can reach from one, when in the considered *region* no influence on the driven distance can be found, up to the number of distances classes we started with. It is not recommended to perform a further split if it is not essential for the correctness of the results. Otherwise, each time a distance class is selected in the simulation, a check which detour factor has to be applied is necessary first.
After the correct mean and standard deviation are found, the chosen distance class is translated like it is described in the next section in order to simulate a suitable linear distance. For our examination of the routes in Germany we prefer the utilization of an individual circuity factor for each distance class with lower limit larger than 2km or 5km. The left-over classes are combined and a single mean and standard deviation are calculated. Of course, other factors not considered here can also have an influence on the detours. In a more detailed analysis also combinations of them should be examined. We tried to fit Gaussian mixture models to the transformed data and conducted regression calculations based on two factors each. We gained no significant improvement on the dependence already found. In the simulation of usage models defects in the final outcomes also arise from other data employed. Thus, the obtained results are of sufficient accuracy for our purposes even if they are not optimal.

5.2.2. Translation of distance classes employing mean and standard deviation

The minimal required input for the translation of distance classes consists of a mean value and the corresponding standard deviation. Even though we have shown in the last section that the assumption of a single value is not correct generally, we now work with an overall mean for all classes in the following in order to simplify the notation, i.e. we focus on one of the components. The required values are also provided in [18] for inter city distances for instance. They can be used directly, especially in the preconditioning iteration when the true values are not known yet, but the options are limited to basic transformations.
From the last chapters we know that the circuity factor is defined as

$$CF = \frac{\text{Driven distance}}{\text{Linear distance}}. \tag{5.2}$$

Since our goal is to determine the linear distances needed in the simulation, we use this formula as

$$\text{Linear distance} = \frac{\text{Driven distance}}{CF}. \tag{5.3}$$

The distance classes reported in mobility statistics shall then be transformed by just recomputing the lower and upper bounds. If only the mean value is inserted as CF in equation (5.3), the variation in the data is neglected. In order to overcome this problem, the computed unknown circuity factors are taken as realizations of a random variable X and Chebyshev's inequality

$$P(|X - E(X)| \geq k\sigma_X) \leq \frac{1}{k^2} \; \forall k > 0, \tag{5.4}$$

compare [35] or [25] for instance, is considered. Here, σ_X represents the standard deviation of X, $E(X)$ its expectation. We try to find some realistic bounds which proceed to limits of the linear distances. Therefore, the given sample mean and sample standard deviations have to be transformed to estimates of expectation and true standard deviation. Unfortunately, figures 5.7(a) in the sections before unmistakably showed that the easiest assumption of an underlying normal distribution avoiding extensive parameter transformations is not realistic. CFs have a natural lower bound of one and their distribution is far from symmetric. However, we stated that the true distribution of all circuity factors follows some mixture distribution with components distinguished by one or more different influencing factors. When using the values given in [18], we assume that we are limited to one homogeneous component, due to the restriction to inter-city connections. Furthermore, we presume that its distribution function has a similar shape and can be approximated well with a shifted log-normal distribution. Therefor a new random variable $Y = X - 1$ is introduced. Obviously, Y possesses the same standard deviation as X, $\sigma_Y = \sigma_X$, its mean value can easily be computed as $\overline{Y} = \overline{X} - 1$. Due to

the fact that only intercity distances are considered, all circuity factors obtained in [18] are assumed to be larger than one. This lower bound can only be attained for completely straight routes without any curves. Hence, the requirement $Y > 0$ necessary for the log-normal distribution is also fulfilled. The goal is now to compute the expectation of Y, transform back to X and apply Chebyshev's inequality. According to [42, page 143] or [47, 136], the expected value of Y following a two-parameter log-normal distribution, $Y \sim LN(\mu, \sigma^2)$, is $E(Y) = \exp\left(\mu + \frac{1}{2}\sigma^2\right)$. Its variance reads $var(Y) = (\exp(\sigma^2) - 1)\exp(2\mu + \sigma^2)$. The population mean and variance can be used as estimators for those quantities. Since $Y = X - 1$, Chebyshev's inequality yields, conditioning on $\overline{Y}$ given,

$$\begin{aligned} \frac{1}{k^2} &\geq P(|Y - E(Y)| \geq k\sigma) \\ &\approx P(|Y - \overline{Y}| \geq k\sigma) \\ &= P(|X - 1 - (\overline{X} - 1)| \geq k\sigma) \\ &= P(|X - \overline{X}| \geq k\sigma). \end{aligned} \tag{5.5}$$

Hence, the next goal is to determine a suitable value for k. Due to the definition of the detour factor and the relation pointed out in equation (5.3), small changes in the values for the CFs lead to significant changes in the computed linear distance. Consequently, a higher probability that the true factor lies in some bounded, but large, interval causes the linear distance to vary drastically.
The condition in the probability of (5.5) gives an interval for the true detour factor,

$$CF \in \left(\overline{CF} - k\sigma, \overline{CF} + k\sigma\right), \tag{5.6}$$

for different values for k. The class limits for driven distances given in [38] shall then be transformed to limits for linear distances. In doing so, as much suitable values as possible should be included which means that resulting left bounds shall be lessened, upper limits shall be raised compared to the use of the mean value. Exploiting the inverse proportionality between linear distance and detour factor, the right limit in (5.6) is used to determine the new minimum, the lower limit is applied for the maximum of the linear distance. Table 5.1 shows the transformation of the limits of the distance classes in the MiD data given in [38] using the smallest multiples of the standard deviation. Rows "+" and "-" indicate the difference in computation.
Unfortunately, only the values for $k \in 0, 1$ are reasonable. In particular, we obtain a change of sign in the subtraction case. Here, $\overline{CF} - k\sigma < 0$ for $k > k_{crit-} = 1.3895$ and the formula for the upper bound turns into one for a lower one. Thus, using this approach, no information on the maximum linear values is given anymore for growing k. The probability for the detour factors to lie in a $k_{crit-}\sigma$ neighborhood around the mean is larger than 0.482 by Chebyshev's inequality. This is of course a more valuable information than that for inserting $k = 1$, but yet not satisfying.

k	Limits [km]	1	2	5	10	20	35	50	80
0	Both	0.76	1.52	3.79	7.58	15.15	26.52	37.88	60.61
1	+	0.44	0.88	2.20	4.41	8.81	15.42	22.03	35.24
	-	2.70	5.41	13.51	27.03	54.05	94.59	135.14	216.22
2	+	0.31	0.62	1.55	3.11	6.21	10.87	15.53	24.84
	-	-1.72	-3.45	-8.62	-17.24	-34.48	-60.34	-86.21	-137.93
3	+	0.24	0.48	1.20	2.40	4.80	8.39	12.00	19.18
	-	-0.65	-1.31	-3.27	-6.54	-13.07	-22.88	-32.68	-52.29
4	+	0.20	0.40	0.98	1.95	3.91	6.84	9.77	15.63
	-	-0.40	-0.81	-2.02	-4.03	-8.06	-14.11	-20.16	-32.26

Table 5.1.: Easy transformation of interval limits of driven distances given in [38] for different factors applying mean and standard deviation for Germany reported in [18]

What is still not exploited is the circumstance that linear distances cannot be larger than the driven ones. This corresponds to the condition that the detour factor is larger or equal than one. Accordingly, the following relation can be derived:

$$\begin{aligned} & \frac{1}{k^2} \geq P\left(|CF - \overline{CF}| \geq k\sigma\right) \\ \Leftrightarrow \quad & 1 - \frac{1}{k^2} \leq P\left(|CF - \overline{CF}| < k\sigma\right) \\ & = P\left(CF \in \left[\overline{CF} - k\sigma, \overline{CF} + k\sigma\right]\right) \\ & = P\left(CF \in \left[\overline{CF} - k\sigma, 1\right)\right) + P\left(\in \left[1, \overline{CF} + k\sigma\right]\right) \\ & = P\left(CF \in \left[1, \overline{CF} + k\sigma\right]\right). \end{aligned} \tag{5.7}$$

The partitioning of the interval is feasible for all k larger than some critical value $k_{crit1} = \frac{\overline{CF}-1}{\sigma}$. In our example, using the values given in [18] for Germany, $k_{crit1} = 0.3368$ which is smaller than the critical value found before. Hence, the expansion of the intervals for $k = 1$ is also directly prevented when the upper limits of the driven distance are used for the linear distances as well if $k > k_{crit1}$. As a results we get feasible lower and natural upper limits. For a single distance class it is valid to take these transformations. An optimal k ensuring a claimed probability can be computed easily from (5.7).
However, considering the complete classification, the resulting bins for the linear distances overlap. Taking another look at table 5.1, the lower interval limits are still given in the "+"-row of each k, the upper bound equals the column heading. Naturally, the lower bounds decrease for growing k, accordingly the intersection between classes increases. In particular, single values might not belong only to two classes but to even more. A linear distance of 0.99 for instance lies in bins 1

and 2 for $k \leq 3$ but additionally in class 3 for $k = 4$. The gain in confidence for a detour factor to lie in the specific interval suffers from inaccuracy in simulated proportions.
Accordingly, the probabilities for the distance classes have to be adapted. After the reasonable selection of k, the limits are translated and the overlap between the distance classes can be computed. Since the original classes are distinct, the resulting intervals can be assigned to the corresponding ones. Then, the size of the intersection and the complete length can be used to calculate the share of overall proportion that has to be redistributed over the new classes. As an example, we take the two short intervals $I_1 = [0, 1)$ and $I_2 = [1, 2)$. We assume that $k = 2$ which gives the partitioning in $\tilde{I}_1 = [0, 1)$ and $\tilde{I}_2 = [0.88, 2)$ for the linear distances during the simulation. Hence, the intersection $I = [0.88, 1)$ obtains a higher probability than expected. If we assume a uniform distribution inside the class limits, we get

$$P(u \in I) = P(I_1) \cdot P(I|I_1) + P(I_2) \cdot P(I|I_2) = 0.12(P(I_1) + P(I_2)) \tag{5.8}$$

with distinct intervals, the second summand was skipped. The probabilities for $\tilde{I}_1 \setminus I$, $\tilde{I}_2 \setminus I$ and I do not sum up to one anymore but overshoot. Of course, depending on the considered *region*, this might not always propagate to the final distribution for the driven distances after the simulation. However, for the multiple runs we performed for different usage models, we always got a remarkable difference to the true distribution.
A possible solution to this problem is the adaption of the probabilities for the single classes used in the random number generator. The overvalued volume has to be extracted. Therefore, one could introduce the intersections as new intervals having a reduced share of the joint proportions or the "old" intervals have to be reduced accordingly. However, more knowledge on the distribution inside the single intervals is required to find a suitable improvement. An optimal solution is not expected to exist since it depends on the underlying road network in combination with the home locations chosen. Some simple tests for $k = 2$ did not succeed since an intended reduction of the proportion of large classes only moved importance to the two next lower classes. Others also decreased their frequency instead of increasing it. An extensive further search exceeds the scope of this work. We accept the methods explained above and rather use different detour factors for the single intervals to enhance the accuracy.

5.3. Influence of the restart of simulation

During the simulation of commuter models sometimes combinations of values are created that do not match. In the common case a chosen distance class is not feasible for a home location. Let us consider the simple situation that a commuter is residing in some rural area, in a small village for instance. There, the density

of working places as well as shopping facilities is usually low, in a certain distance often no feasible target points exists. Thus, there exist classes in the distance distribution that are not suited for this home. Nevertheless, they can be selected. They can even be valid for people living in an other part of the neighborhood. Especially if small communes are aggregated in one ROI only. Hence, the choice of that distance bounds is just a problem for that individual commuter, not a general one, and cannot be removed from the random selection.
Such a situation can be solved in two ways. First of all, the distance is simulated anew and the home location and all other trips determined so far are kept. Secondly, the complete iteration is restarted. This requires more effort than the former approach since feasible parts of the trip are dropped here and have to be created again. Nevertheless, the simulation performs more stable in that case. It could happen that, due to errors in the map data for instance, none of the distribution classes fits at all. In order to prevent too large trips, the distance classes are bounded such that the largest possible value might still be too small. Particularly this can happen if different distributions depending on the trip type are applied. Then a route towards a target point is created but there is no feasible destination lying in acceptable distance. Hence, the specified commuter cannot exist and has to be determined correctly. In the first approach, some back-propagation could be caused, stopping and restarting somewhere inside the complete simulation procedure. In the second approach, the behavior can be supervised better.
It has to be checked, if the prescribed proportions and distributions are correctly reproduced after the restart of the simulation. We therefore log some intermediate values and evaluate the created kml-files. The influence of the processing with VMC® is skipped here since it is not one the main parts of the usage modeling. It is obvious that missing bridges or newly constructed roads for instance could affect the results. The following review is conducted on ten runs for the simulation of 1,000 independent commuters in Germany each.
We start with the distribution of home locations. These form the center of the driving pattern and have to be selected properly. For the 10 series, we obtain a mean value of 2.2 restarts. For that calculation just the initial and resulting places of residence is compared. In all cases only one recalculation was necessary. Thus, for the computation of driving patterns of 1,000 persons 1,002.2 single iterations of the complete method are required on average, not producing much additional effort. The small number of changed home locations directly verifies that the obtained vehicle distribution, which was shown to be reflected well in chapter 4.2 and 5.1, is not destroyed.
The same holds for the proportions of the different commuter route types. The obtained frequencies of the three different types are depicted in figure 5.11. The dotted histogram indicates the prescribed discrete distribution. Obviously, the fractions are reflected well. This visual result can be verified statistically by the computation of confidence intervals for the proportions like it is described in [23] and [24]. In our case, the sample size of only 10 for the comparison is not large enough to apply the central-limit theorem enabling the use of the nor-

mal distribution. However, if we separate the 10,000 single commuters in 100 groups of 100 individuals each, we can calculate the three confidence intervals. Of course, some samples can be excluded from that process forming a test set to check the accuracy of the simulation. The obtained results were quite good.

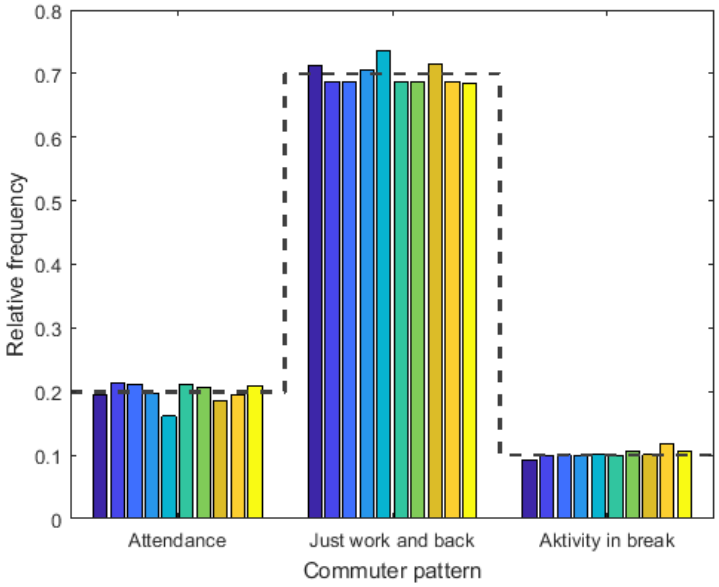


Figure 5.11.: Frequency distribution obtained for the simulation of 1,000 commuters in each run.

Due to the problem of determining the correct settings for the translation of distance classes explained in chapter 5.2.2, the obtained driven distances to not fit the prescribed ones that good. The application of an average detour factor for all distances led to reduced proportions of large distances on the linear case but yielded a huge amount of long trips after the routing. It was examined that the change from $k = 1$ to $k = 2$ produced the expected positive effect on the result, yet it was not optimal. Thus, if suitable translations between the intervals are found, also these outcomes should be enhanced.
Of course, some discrepancies are always expected since the selection of participants of the survey and especially the infrastructure and road network around their place of residence have an influence that cannot be ignored. Additionally, for our comparisons we employed the distributions freely available online in the report of the MiD2008 [38] where no detailed information on the fraction of trips conducted by car is contained in the applied statistics. Hence, also this uncertainty might derange the outcomes. The small number of required restarts shown before indicates that once a suitable distribution is used, it will be reflected well by the algorithm.

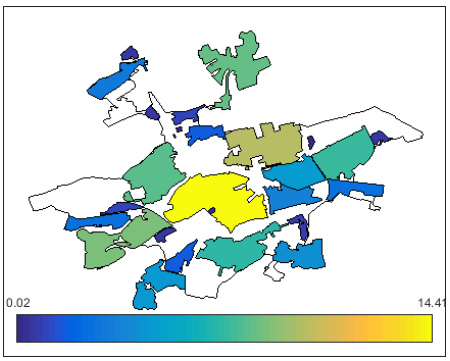
6. Application examples

In the last part of chapter 2 we already repeated the example how the usage model for light-duty commercial vehicles can be combined with measurement campaign to estimate reliable pseudo-damage values presented in [22]. There, the simulation played a minor role and we concentrated on the determination of attributes related to the driven routes. Here, we want to sketch the single steps during the simulation of a usage model. We therefore consider a typical commuter living in Kaiserslautern as an example. We depict in detail how the different types of data introduced in chapter 4 are employed. We finish with the computation of speed profiles for taxis. As an example, we compare the influence of the driving style of an aggressive with that of a careful driver.

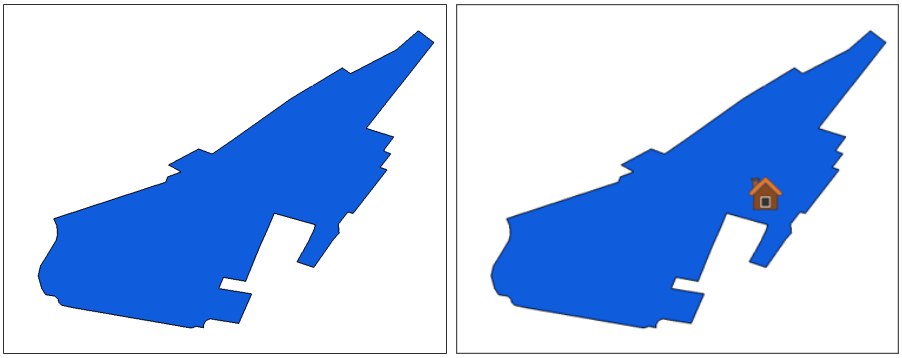
6.1. Simulation of representative routes for commuters

In order to illustrate the steps conducted in the simulation of usage models, the example of a commuter living in Kaiserslautern given in [21] is reproduced adding some further details. First, the model components have to be initialized.
It is assumed that also the workplace shall be located there, hence *region* is set to Kaiserslautern including suburban municipalities. Adequate statistics for all required distance distributions are available and already transformed using detour factors. P is chosen as the base pattern without special regular trip, D can then be determined easily containing three different types of distance classes. One for the way between home and work and two for the leisure time activities. The first more involving step is the simulation of B. Figure 6.1 depicts the selection of a home location in three steps. At the beginning, all residential areas of Kaiserslautern are rated by their size. One of them is selected randomly and the home location is placed there afterwards applying the methods in chapter 4.1.3. The central location B is indicated by a small house in the following.

In the next step, the first entry of D is used to search a workplace. Here, only ROIs are considered as an example. Figure 6.2 shortly sketches the procedure. First, all feasible areas are intersected with the requested distance class represented by a circular ring. Afterwards a specific location is chosen, flagged with the high-rise building. Since the simplest commuter pattern has been chosen, no attendance

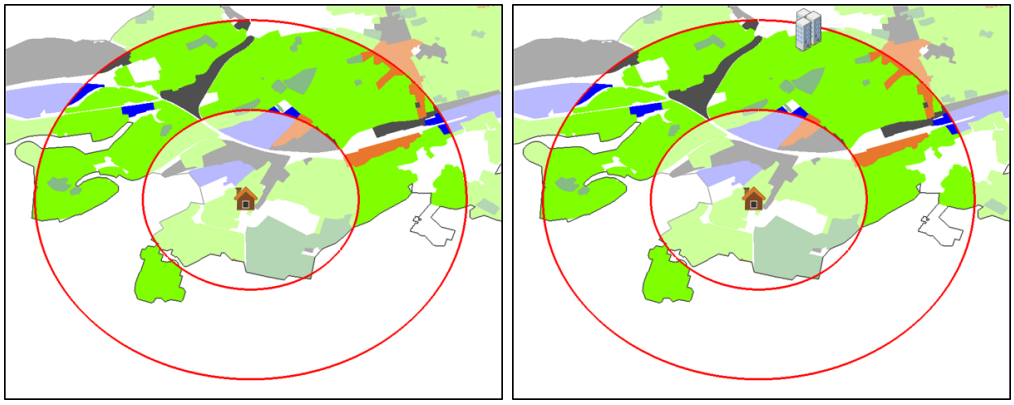


(a) Residential areas in Kaiserslautern



(b) Selection of a single neighborhood (c) Choice of coordinates for B

Figure 6.1.: Choice of B conducted in three steps; illustration embedding [27].



(a) Determination of suitable areas (b) Selection of a work location

Figure 6.2.: Choice of a workplace in two steps; own illustration using [27, 28].

or break activities have to be simulated. Thus, the stopovers can be determined next. It is prescribed that exactly two additional destinations are approached over the week. Then, feasible target points have to be determined. If no further restrictions are set, all POIs located in Kaiserslautern are used as a basis, see

figure 6.3. By chance, the same distance class has been chosen for both leisure

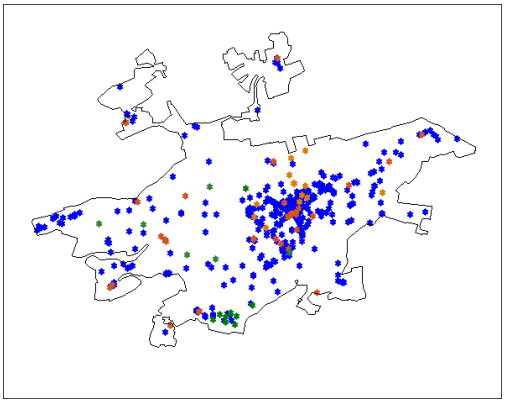
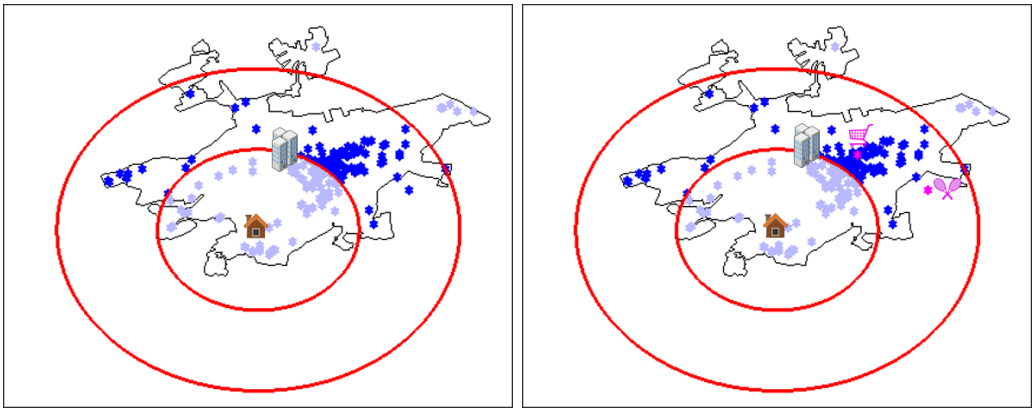


Figure 6.3.: Feasible target points for stopovers in Kaiserslautern

time activities. Hence, figure 6.4 summarizes the selection of both. Usually, the described steps have to be repeated with the different entries of D. The remaining, correctly classified, POIs are checked if the fulfill the requested distance conditions. Afterwards, specific ones are selected randomly. In that process no rating is conducted, all points have equal probability. The two resulting target points are indicated in the right graphic. One destination is classified as shopping facility, the other is placed at a sports center.



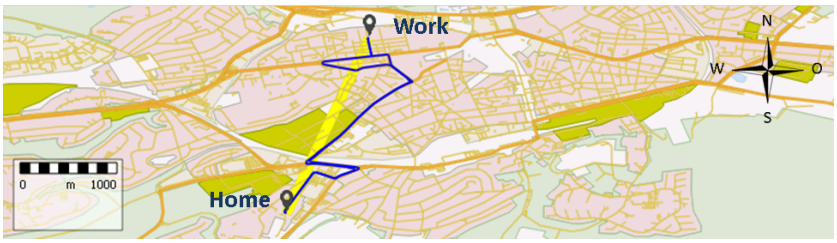
(a) Determination of suitable POIs (b) Selection of two stopovers

Figure 6.4.: Choice of stopovers in two steps, own illustration embedding [27, 28, 29, 30].

The last required input of the route assembling is the scheduling of the leisure time activities. For these the days are chosen first. In the simple model it is assumed that all days have equal probability, thus a basis integer random number generator is used to independently select two numbers between one and five. In the example given in [21], the number three is chosen twice. This means both trips are performed on Wednesday.

Next, the chronological order on this day has to be determined. Again, two binary random numbers are generated. In this process the frequency distribution provided in MP is used. In the example one additional destination is approached as stopover on the way home, the other is visited on a new trip.
The routes on the five days can be split in two groups. On Monday, Tuesday, Thursday and Friday the simulated commuter only travels from home to his work location and returns directly afterwards. They are visualized in figure 6.5. The upper picture shows the two locations on the map. They are connected linearly. The lower screenshot depicts the result after the routing algorithm of VMC® has been applied. First of all, the proposed coordinates are projected to the nearest suitable roads. Afterwards, a path through the road network is determined.
Graphic 6.6 shows the trips performed on Wednesday. In addition to the already known locations, also the leisure time activities are inserted. All ways are summarized on one map but the two trip chains starting from the home location can be identified easily. Again, the resulting routes can be segmented and analyzed

(a) Home and work location connected linearly



(b) Home and work connected on road network

Figure 6.5.: Home and work location projected on map, connected linearly and by feasible route, compare [21]

by VMC®. The results provided in [21] are repeated in tables 6.1 and 6.2. Obviously, no motorway is passed since the simulated commuter only travels inside the city and Kaiserslautern does not posses an urban motorway. Additionally, only the extra trip is performed on rural roads and has only a minor share on the overall driven distance during week. Due to the topography of the city the slopes are not that severe and most roads traveled are rather flat.

Figure 6.6.: More complex route on Wednesday, destination in eastern part of city is visited on extra trip starting from home.

Motorway	Rural A	Rural B+C	Urban A	Urban B+C	Other
0%	3.1%	6.6%	21.6%	66.3%	2.3%

Table 6.1.: Split of all routes on five days according to road type, see [21]

Slope class	0-3%	3-6%	6-9%	9-12%	12-15%	>15%
Proportions	83.2%	14.5%	1.8%	0.2%	0.3%	0.0%

Table 6.2.: Split of all routes on five days according to slope classes, compare [21]

6.2. Using created routes: Speed profiles for taxis

Like in the last section, we choose the base location of the simulated taxi to be settled in Kaiserslautern. Due to the different trip types and distance distribution resulting from KiD data [26], the car also leaves the city, but it is forced to return back after each tour. Again, the kml-files created during the simulation of the taxi model are routed with VMC® to find the streets traveled. We always prefer the shortest distance, since often taxi prices are composed of a basic charge and costs per kilometer. Hence, large detours are not allowed. In the end we obtain the routes for seven days. For these different speed profiles are calculated. In that process we employ each trip twice. On one hand we assume that the driver is careful and rather respects the legal speed limits. On the other hand, a more aggressive behavior is simulated.
For the comparison of the results we concentrate on the single trip between two suburbs shown in figure 6.7. It is traveled from east to west. Hence, it starts in some residential area, then a major road is reached shortly after and it finishes on small roads in an industrial area. It has a length of nearly 3km. The route

Figure 6.7.: Taxi trip used for comparisons. The route is conducted from east to west.

itself can be analyzed for different attributes. Figure 6.8 for instance presents its slope and curvature. The sharp turns required to enter and leave the major road can be detected easily in the curvature. These signals along the route have an influence on the velocity of a vehicle driving there. However, the impact of the different styles can be easily detected in figure 6.9. Both curves are created with VMC® and summarized in one plot to facilitate the comparison. As expected, the velocity of the careful driver is nearly always lower than that of the aggressive motorist. The latter also breaks and accelerates more often. The resulting forces in direction of and orthogonal to the driving direction are presented in figure 6.10. Especially the latter have a significantly larger amplitude for the aggressive driver, but also those measured in the driving direction differ. Here, the larger accelerations are visible. VMC® also offers to draw graphics for lateral and longitudinal acceleration, though these shall be skipped here since they offer no essentially new information. All types of analysis can be conducted and corresponding plots can be created for a large number of simulated taxis for all days and trips. The acceleration or the forces can for instance be divided into different classes. Based on those, the two motorists can be well compared. Some additional load data analysis can be applied to examine the influence of the driving style on the damage of a vehicle.

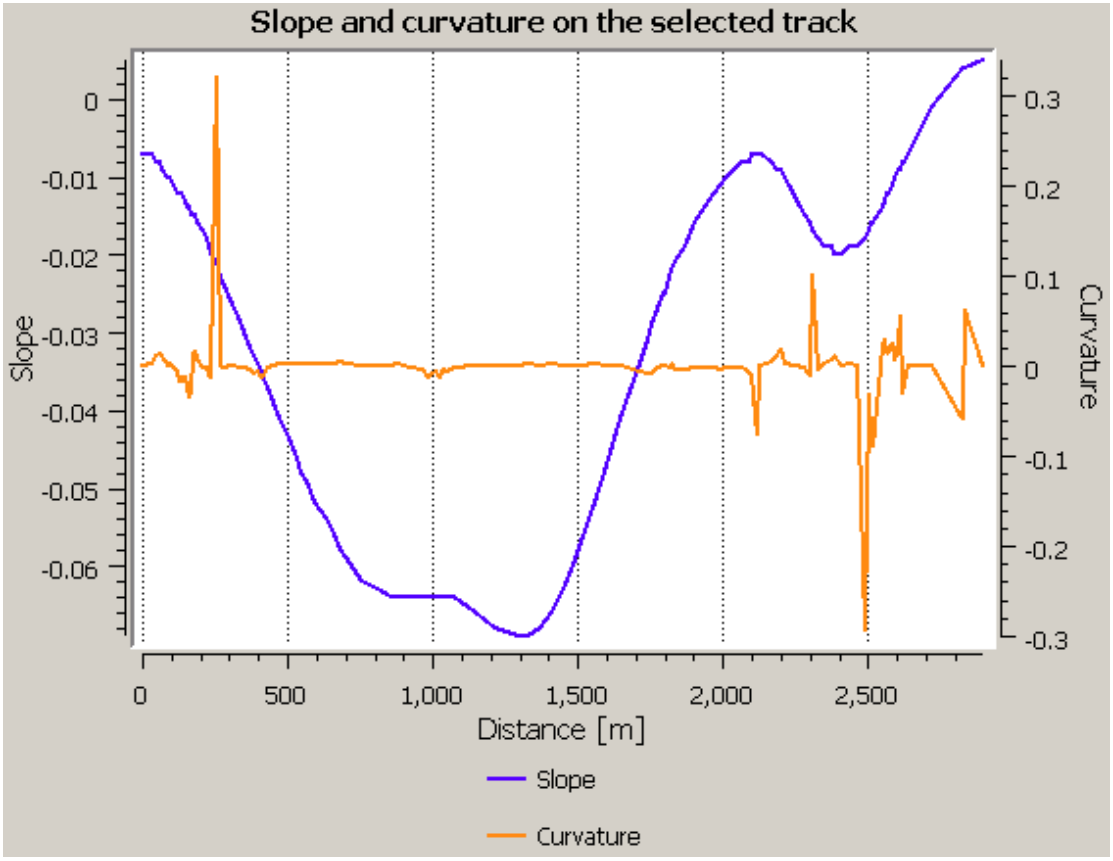


Figure 6.8.: Slope and curvature measured on the roads traveled.

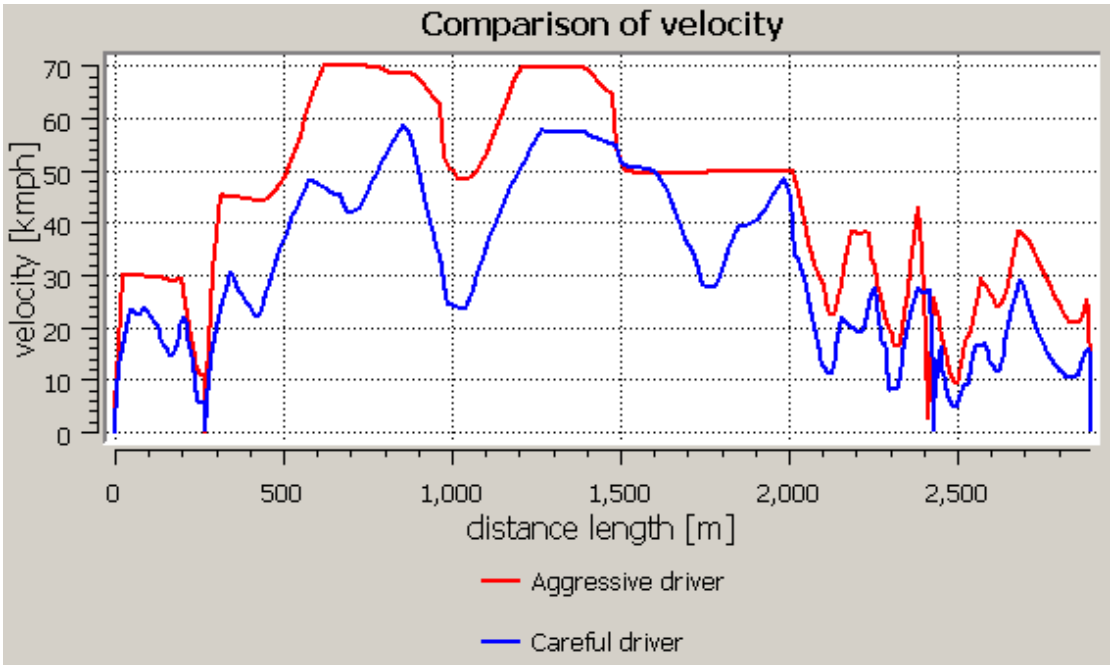
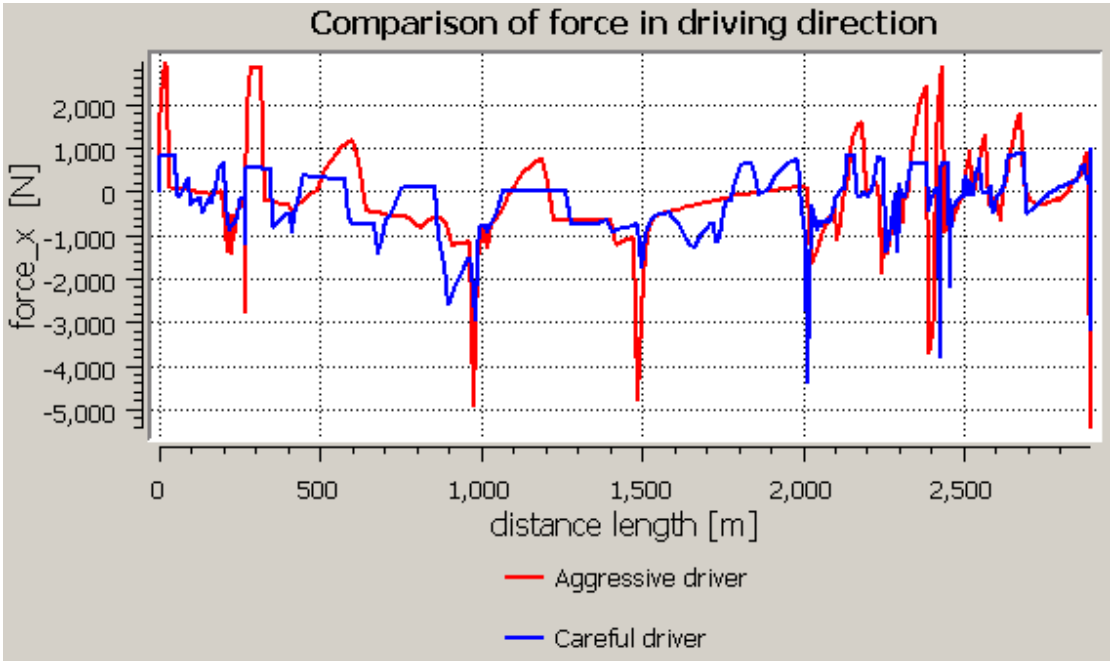
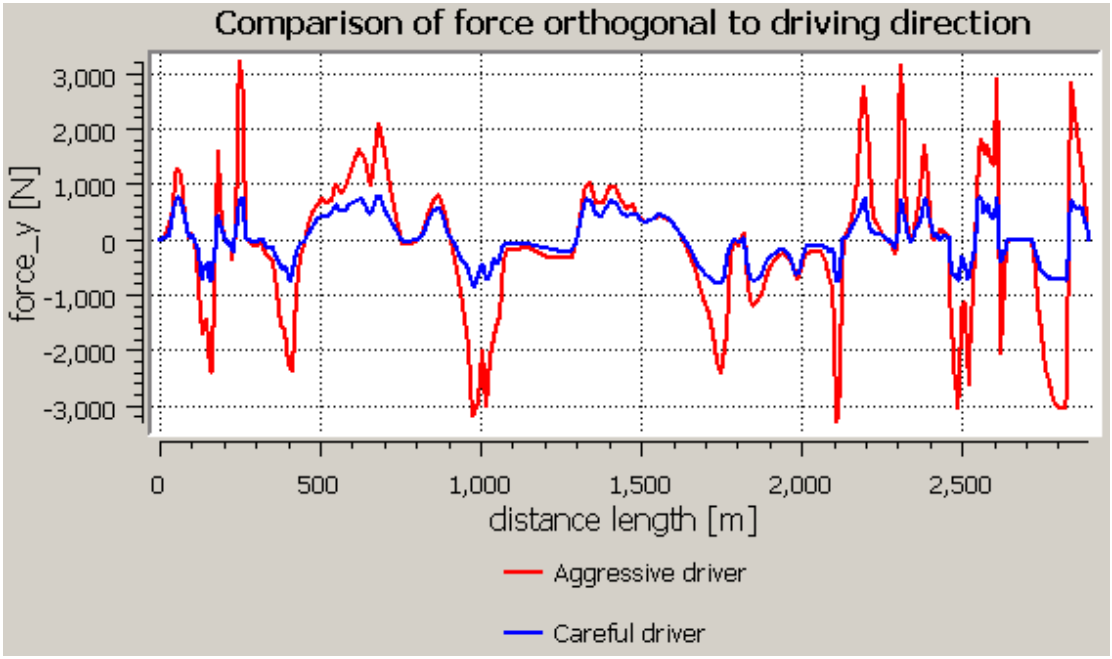


Figure 6.9.: Comparison of velocity of aggressive and careful driver.



(a) Force in direction of the driving direction.



(b) Force orthogonal to the driving direction.

Figure 6.10.: Comparison of forces for aggressive and careful driver.

7. Summary and further prospects

In this work the concept of usage models and subproblems that have to be solved was developed and discussed. We started with the mathematical description of models for different user types. Afterwards, we considered the different kinds of data required to simulate reliable routes. The algorithm applied for the creating of single routes was introduced. Especially the determination of a correct population distribution and the selection of specific home locations was demonstrated in detail. We showed, that the population counts included in the OSM data are of sufficient quality and currentness to rely on. We also proved that population data is a good indicator for the number of vehicles registered in an administrative unit. Thus, the vehicle distribution is reflected well and we can simulate home locations of individuals without providing additional data. The VMC® database is sufficient.
The computation of detour factors was not that successful. Alternating the simulation of the usage model, routing of the created trips with VMC® and computing the resulting circuity factors could be performed to enhance the mean values given in [18], but it is too costly. Unfortunately, the determination of influencing factors based on a single iteration did not yield satisfying results. We expect that the CFs follow a mixture of log-normal distributions, but we are not able to prove this assumption. Quantile-quantile and normal probability plots for the detour $CF-1$ show some linear behavior for the middle part of the data, but trimming lower and higher values does not cause positive test results. Additionally, the intervals for linear distances required in the simulation overlap and produce wrong shares of some classes in the final results. An optimal value for k as well as a method to distribute the augmented volume has not been found, yet.
Hence, future work has to be invested in this topic. The combination of multiple influencing factors has to be considered. Especially some characteristics not at hand at the moment, like the accessibility of motorways or the fraction of motorway kilometers on the complete road network in an administrative unit should be investigated. After more suitable CFs are determined, also the effect of the routing algorithm applied can be examined. There should be a difference between shortest and fastest ways.
Further tasks consider the usage models themselves. That of light-duty commercial vehicles should be adapted for heavy-duty trucks. In the commuter model,

leisure time activities on the weekend as well as vacation trips shall be included. Professional trips conducted in private cars should be checked for relevance. Additionally, socio-economic factors splitting population in workers, non-workers and retired people should be considered.
Population distribution based on communes should be check for availability. Therefore, the population counts on this lower administrative level have to be of good quality.
According to the traffic surveys, iterative proportional fitting can be applied to filter only relevant distances covered by cars.

This list of problems and extensions is rather long and might create the impression that the presented method of automatic usage modeling for automotive applications could not be working reliably, yet. Different use cases however already showed that the results are quite good to reflect complete populations. The simulated routes can be evaluated with VMC® and then used for various purposes. They can be combined with the results of measurement campaigns providing the settings for the usage simulation with the U·Sim software package. Hence, suitable pseudo-damage values can be computed. Additionally, they can be applied for all types of evaluations in VMC® requiring routes. Different regions or driving styles can be compared easily. In that process, the behavior of a complete population is reproduced.
Summarizing, usage models form a good basis to estimate the loads acting during typical vehicle use.

Bibliography

[1] Federal Ministry of Transport and Digital Infrastructure. http://www.bmvi.de.

[2] German Aerospace Center (DLR). http://www.dlr.de.

[3] GPS Data Team. See https://www.gps-data-team.com.

[4] infas Institute for Applied Social Sciences. http://www.infas.eu.

[5] Karte in OpenStreetMap. See https://www.openstreetmap.de/karte.html, This data is made available under the Open Database License: http://opendatacommons.org/licenses/odbl/1.0/. Any rights in individual contents of the database are licensed under the Database Contents License: http://opendatacommons.org/licenses/dbcl/1.0/.

[6] Mobilität in Deutschland 2008 (MiD 2008). See www.mobilitaet-in-deutschland.de.

[7] MultiPolygon. See https://msdn.microsoft.com/de-de/library/bb964739.aspx.

[8] Open street map wiki, key:population. See https://wiki.openstreetmap.org/wiki/Key:population, retrieved May 31st, 2017, last modified on November 21st, 2015.

[9] OpenStreetMap. See www.openstreetmap.org.

[10] OpenStreetMap - Deutschland, FAQs. See https://www.openstreetmap.de/faq.html\#was_ist_osm.

[11] OpenStreetMap Wiki, join the community. See http://wiki.openstreetmap.org/wiki/Join_the_community, retrieved December 1st, 2015, last modified on September 3rd, 2015.

[12] POIbase. See https://www.poibase.com/en.

[13] POIplaza. See poiplaza.com.

[14] Polygon. See https://docs.microsoft.com/de-de/sql/relational-databases/spatial/polygon.

[15] Polygon in PostgreSQL. See https://www.postgresql.org/docs/9.4/static/datatype-geometric.html.

[16] taginfo - de:amtlicher_gemeindeschluessel. See https://taginfo.openstreetmap.org/keys/de:amtlicher_gemeindeschluessel.

[17] taginfo - opengeodb:community_identification_number. See https://taginfo.openstreetmap.org/keys/openGeoDB:community_identification_number.

[18] R. H. Ballou, H. Rahardja, and N. Sakai. Selected country circuity factors for road travel distance estimation. *Transportation Research Part A: Policy and Practice*, 36(9):843–848, 2002.

[19] R. J. Beckman, K. A. Baggerly, and M. D. McKay. Creating synthetic baseline populations. *Transportation Research Part A: Policy and Practice*, 30(6):415–429, 1996.

[20] C. Bhat, J. Guo, S. Srinivasan, and A. Sivakumar. Comprehensive Econometric Microsimulator for Daily Activity-Travel Patterns. *Transportation Research Record*, 1894:57–66, 2004.

[21] C. Biedinger and S. Feth. Usage modeling of commuters on basis of geographical data for vehicle engineering. In *Young Researchers Symposium 2016 (YRS 2016)*, Stuttgart, 2016. Fraunhofer Verlag.

[22] C. Biedinger, T. Weyh, A. Opalinski, and M. Wagner. Simulation of customer-specific vehicle usage. In K. Berns, editor, *Commercial vehicle technology 2016*. Shaker Verlag, Aachen, 2016.

[23] M. Birkin and M. Clarke. Synthesis—A Synthetic Spatial Information System for Urban and Regional Analysis: Methods and Examples. *Environment and Planning A*, 20(12):1645–1671, 1988.

[24] H. M. Blalock. *Social statistics*. McGraw-Hill series in sociology. McGraw-Hill Kogakusha, Tokyo, 2. ed., internat. student ed. edition, 1972.

[25] J. Bortz. *Lehrbuch der Statistik: Für Sozialwissenschaftler*. Springer, Berlin [etc.], mit 69 abbn. und 213 tabn edition, 1977.

[26] Bundesministerium für Verkehr und digitale Infrastruktur. Kraftfahrzeugverkehr in Deutschland 2010 (KiD2010), 2012. www.kid2010.de.

[27] Clker-Free-Vector-Images, https://pixabay.com/users/Clker-Free-Vector-Images-3736/?utm_source=link-attribution&utm_medium=referral&utm_campaign=image&utm_content=295214. Downloaded from https://pixabay.com/de/vectors/geb%C3%A4ude-haus-home-bau-immobilien-295214, color was changed.

[28] Clker-Free-Vector-Images, https://pixabay.com/users/Clker-Free-Vector-Images-3736/?utm_source=link-attribution&utm_medium=referral&utm_campaign=image&utm_content=295511. Downloaded from https://pixabay.com/de/vectors/geb%C3%A4ude-hochhaus-h%C3%A4user-immobilien-295511.

[29] Clker-Free-Vector-Images, https://pixabay.com/users/Clker-Free-Vector-Images-3736/?utm_source=link-attribution&utm_medium=referral&utm_campaign=image&utm_content=304843. Downloaded from https://pixabay.com/vectors/shopping-cart-caddy-shopping-trolley-304843, color was changed.

[30] Clker-Free-Vector-Images, https://pixabay.com/users/Clker-Free-Vector-Images-3736/?utm_source=link-attribution&utm_medium=referral&utm_campaign=image&utm_content=310390. Downloaded from https://pixabay.com/vectors/racquets-rackets-tennis-sport-310390, color was changed.

[31] C. L. Eicher and C. A. Brewer. Dasymetric Mapping and Areal Interpolation: Implementation and Evaluation. *Cartography and Geographic Information Science*, 28(2):125–138, 2001.

[32] D. J. Giacomin and D. M. Levinson. Road network circuity in metropolitan areas. *Environment and Planning B: Planning and Design*, 42(6):1040–1053, 2015.

[33] I. N. Gregory. The accuracy of areal interpolation techniques: Standardising 19th and 20th century census data to allow long-term comparisons. *Computers, Environment and Urban Systems*, 26(4):293–314, 2002.

[34] S. I. Hay, A. M. Noor, A. Nelson, and A. J. Tatem. The accuracy of human population maps for public health application. *Tropical medicine & international health : TM & IH*, 10(10):1073–1086, 2005.

[35] H.-D. Hippmann. *Statistik: Praxisbezogenes Lehrbuch mit Beispielen.* Schäffer-Poeschel, Stuttgart, 4., überarb. aufl. edition, 2007.

[36] infas Institut für angewandte Sozialwissenschaft GmbH and Deutsches Zentrum für Luft- und Raumfahrt e.V. Institut für Verkehrsforschung. Mobilität in Deutschland 2008 Methodenbericht, 2010. www.mobilitaet-in-deutschland.de/pdf/MiD2008_Methodenbericht_I.pdf.

[37] infas Institut für angewandte Sozialwissenschaft GmbH and Deutsches Zentrum für Luft- und Raumfahrt e.V. Institut für Verkehrsforschung. Mobilität in Deutschland 2008 Nutzerhandbuch, 2010. www.mobilitaet-in-deutschland.de/pdf/MiD2008_Nutzerhandbuch.pdf.

[38] infas Institut für angewandte Sozialwissenschaft GmbH and Deutsches Zentrum für Luft- und Raumfahrt e.V. Institut für Verkehrsforschung. Mobilität in Deutschland 2008 Tabellenband, 2010. www.mobilitaet-in-deutschland.de/pdf/MiD2008_Tabellenband.pdf.

[39] IRU-Taxigruppe. Faktensammlung TAXI - flexibel. http://www.bzp.org/Content/RUND_UMS_TAXI/Merkblatt_D_-_Flexibel.pdf.

[40] H. H. Kim. *Intelligent interpolation for population distribution modeling.* Dissertation, University of Georgia, Athens, Georgia, 08/2009.

[41] Kraftfahrt Bundesamt. Fahrzeugzulassungen (FZ): Bestand an Kraftfahrzeugen und Kraftfahrzeuganhängern nach Zulassungsbezirken 1.Januar 2016. 2017. http://www.kba.de/SharedDocs/Publikationen/DE/Statistik/Fahrzeuge/FZ/2016/fz1_2016_pdf.pdf?__blob=publicationFile&v=5.

[42] E. T. Lee. *Statistical methods for survival data analysis.* Wiley series in probability and mathematical statistics : Applied probability and statistics. Wiley, New York, 2. ed. edition, 1992.

[43] D. Levinson and A. El-Geneidy. The minimum circuity frontier and the journey to work. *Regional Science and Urban Economics*, 39(6):732–738, 2009.

[44] S. L. Lohr. *Sampling: Design and analysis.* Duxbury Press, Pacific Grove, 1999.

[45] R. Lovelace, D. Ballas, and M. Watson. A spatial microsimulation approach for the analysis of commuter patterns: From individual to regional levels. *Journal of Transport Geography*, 34:282–296, 2014.

[46] Ministère de l'écologie, du développement durable et de l'énergie (MEDDE), Commissariat Général au Développement Durable (CGDD), Service de l'Observation et des Statistiques (SOeS). Les déplacements locaux un jour de semaine selon les motifs -hors marche à pied, enquête nationale transports et déplacements (entd) 2008, 2010.

[47] D. C. Montgomery and G. C. Runger. *Applied statistics and probability for engineers.* John Wiley, Hoboken, N.J., 5th ed., si version edition, 2011.

[48] C. S. Phibbs and H. S. Luft. Correlation of travel time on roads versus straight line distance. *Medical Care Research and Review*, 52(4):532–542, 1995.

[49] W.-D. Rase. Volume-preserving interpolation of a smooth surface from polygon-related data. *Journal of Geographical Systems*, 3(2):199–213, 2001.

[50] Richerd Darbéra. Taxicab regulation and urban residents' use and perception of taxi services: a survey in eight cities. In *12th World Conference on Transport, Jul 2010, Lisbonne, Portugal.*

[51] B. Schuppar. *Geometrie auf der Kugel: Alltägliche Phänomene rund um Erde und Himmel*. Mathematik Primarstufe und Sekundarstufe I + II. 2017.

[52] M. Speckert, K. Dreßler, M. Lübke, and T. Halfmann. Automatisierte und um Geo-daten angereicherte Auswertung von Messdaten zur Herleitung von Beanspruchungsverteilungen. In *Tagung des DVM-Arbeitskreises Betriebsfestigkeit, Deutscher Verband für Materialforschung und -prüfung 2016 – Potenziale im Zusammenspiel von Versuch*, volume 143, pages 165–180. 2016.

[53] M. Speckert, K. Dreßler, N. Ruf, T. Halfmann, and S. Polanski. The virtual measurement campaign (vmc) – a methodology for geo-referenced description and evaluation of environmental conditions for vehicle loads and energy efficiency. In *3rd Commercial Vehicle Technology Symposium*, pages 88–98. 2014.

[54] M. Speckert, K. Dreßler, N. Ruf, R. Müller, and C. Weber. Customer usage profiles, strength requirements and test schedules in truck engineering. In *1st Commercial Vehicle Technology Symposium*, pages 298–307, 2010.

[55] Statistische Ämter des Bundes und der Länder. Bevölkerungsstand: Durchschnittliche Jahresbevölkerung nach Geschlecht - Jahresdurchschnitt -regionale Tiefe: Kreise und krfr. Städte: Zeitraum: 2015. `https://www.regionalstatistik.de/genesis/online/`.

[56] Statistisches Bundesamt. Gemeindeverzeichnis Gebietsstand: 31.12.2015 (Jahr), July 2017. `https://www.destatis.de/DE/ZahlenFakten/LaenderRegionen/Regionales/Gemeindeverzeichnis/Administrativ/Archiv/GVAuszugJ/31122015_Auszug_GV.html`.

[57] Y. Susilo and R. Kitamura. Analysis of day-to-day variability in an individual's action space: Exploration of 6-week mobidrive travel diary data. *Transportation Research Record: Journal of the Transportation Research Board*, 1902:124–133, 2005.

[58] W. R. Tobler. Smooth pycnophylactic interpolation for geographical regions. *Journal of the American Statistical Association*, 74(367):519–530, 1979.

[59] C. Veness. Movable Type Scripts: Calculate distance, bearing and more between Latitude/Longitude points, January 2015. `www.movable-type.co.uk/scripts/latlong.html`.

[60] T. Weyh, M. Speckert, A. Opalinski, and M. Wagner. Planning a measurement campaign in Eastern Europe using Fraunhofer VMC ("Virtual Measurement Campaign"). In *Commercial Vehicles 2015*, VDI-Berichte, 0083-5560. VDI-Verlag GmbH, Düsseldorf, 2015.

[61] E. H. Yoo, P. C. Kyriakidis, and W. Tobler. Reconstructing population density surfaces from areal data: A comparison of tobler's pycnophylactic interpolation method and area-to-point Kriging. *Geographical Analysis*, 42(1):78–98, 2010.

List of Figures

Figures without reference are created with MATLAB or VMC® on basis of simulation results.
Geo-referenced data visualized is extracted from OpenStreetMap [9]. This data is made available under the Open Database License: `http://opendatacommons.org/licenses/odbl/1.0/`. Any rights in individual contents of the database are licensed under the Database Contents License: `http://opendatacommons.org/licenses/dbcl/1.0/`.

List of Tables

A. Background for simulation of geographical coordinates

In chapters 4.1.3 and 4.1.4 the simulation of specific geographic coordinates within a certain distance to some central point was brought to the issue. There, we did no go into detail how the new point can be determined after the linear distance measured on the surface of the earth and the initial bearing have been selected. The method applied, based on spherical geometry, shall be explained here. Figure A.1 visualizes a possible result on the earth surface. This special case where both points lie on the northern hemisphere and the target point B is in the east of the initial point A is used to demonstrate the basic procedure. The adaption of signs for other configurations is skipped here.

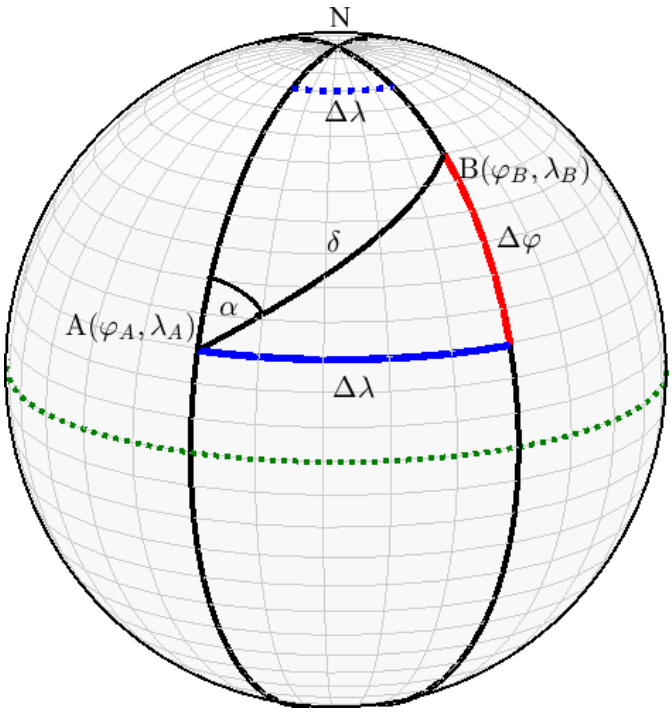


Figure A.1.: Sketch of the situation on the earth's surface. Picture was drawn on basis of [51, fig. 3.1, page 34].

The input variables for the computation are the geographic coordinates (φ_A, λ_A) of A, the initial bearing α and the great circle distance δ. The goal is the calculation

of the position of B, also given by its coordinates (φ_B, λ_B). In order to do this, the colored component-wise differences $\Delta\varphi$ and $\Delta\lambda$ are determined. As already indicated in the figure, the angle $\Delta\lambda$ can also be found in the spherical triangle between A, B and the north pole N. It is extracted in figure A.2. The inscribed length of the edge between A and N results from the fact that meridians have a length of π measured between the two poles and thus $\frac{\pi}{2}$ on one hemisphere. The latitude φ_A has to be subtracted to obtain the true edge length. The same holds for the other edge.

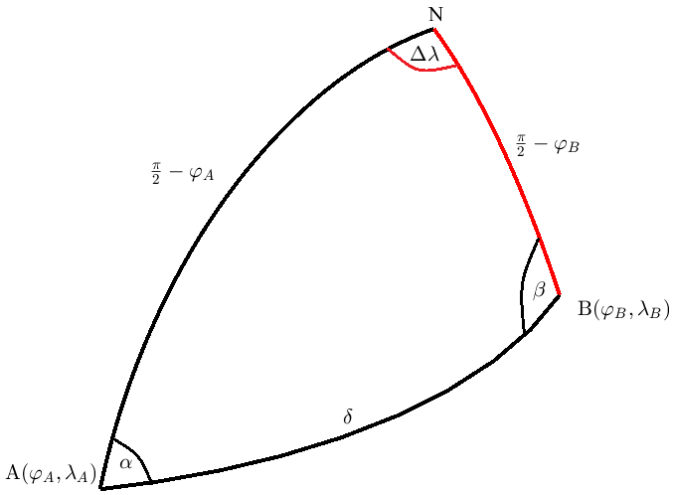


Figure A.2.: Sketch of the spherical triangle, adapted to [51, fig 5.1, page 79].

The spherical version of the cosine rule can be used to compute φ_B. It reads

$$cos\left(\frac{\pi}{2} - \varphi_B\right) = cos\delta \cdot cos\left(\frac{\pi}{2} - \varphi_A\right) + sin\delta \cdot sin\left(\frac{\pi}{2} - \varphi_A\right) \cdot cos\alpha \qquad \text{(A.1)}$$

compare [51, eq. (4.1), page 63]. It can easily be transformed to

$$\begin{aligned} & cos\left(\frac{\pi}{2} - \varphi_B\right) = cos\delta \cdot cos\left(\frac{\pi}{2} - \varphi_A\right) + sin\delta \cdot sin\left(\frac{\pi}{2} - \varphi_A\right) \cdot cos\alpha \qquad \text{(A.2)} \\ \Leftrightarrow \quad & sin\varphi_B = cos\delta \cdot sin\varphi_A + sin\delta \cdot cos\varphi_A \cdot cos\alpha \\ \Leftrightarrow \quad & \varphi_B = arcsin\left(cos\delta \cdot sin\varphi_A + sin\delta \cdot cos\varphi_A \cdot cos\alpha\right). \end{aligned}$$

The computation of a formula for $\Delta\lambda$ is more demanding. It also starts with the cosine rule, this time applied for δ, and is solved for $\Delta\lambda$. In order to simplify the nota-

tion, the terms including $\frac{\pi}{2}$ have been replaced directly.

$$cos\delta = sin\varphi_B \cdot sin\varphi_A + cos\varphi_B \cdot cos\varphi_A \cdot cos\left(\Delta\lambda\right) \tag{A.3}$$

$$\Leftrightarrow \quad cos\left(\Delta\lambda\right) = \frac{cos\delta - sin\varphi_B \cdot sin\varphi_A}{cos\varphi_B \cdot cos\varphi_A}$$

$$\Leftrightarrow \quad cos\left(\Delta\lambda\right) = \frac{\left(sin^2\varphi_A + cos^2\varphi_A\right) cos\delta - sin\varphi_B \cdot sin\varphi_A}{cos\varphi_B \cdot cos\varphi_A}$$

$$\Leftrightarrow \quad cos\left(\Delta\lambda\right) = \frac{cos\varphi_A cos\delta}{cos\varphi_B} + \frac{sin^2\varphi_A \cdot cos\delta - sin\varphi_B \cdot sin\varphi_A}{cos\varphi_B \cdot cos\varphi_A}$$

$$\Leftrightarrow \quad cos\left(\Delta\lambda\right) = \frac{cos\varphi_A cos\delta}{cos\varphi_B} + \frac{sin\varphi_A \cdot sin\delta \cdot \left(-cos\alpha\right)}{cos\varphi_B}$$

$$\Leftrightarrow \quad cos\left(\Delta\lambda\right) = \frac{cos\varphi_A \cdot cos\delta - cos\alpha \cdot sin\varphi_A \cdot sin\delta}{cos\varphi_B}$$

$$\Leftrightarrow \quad cos\left(\Delta\lambda\right) \cdot cos\varphi_B = cos\varphi_A \cdot cos\delta - cos\alpha \cdot sin\varphi_A \cdot sin\delta$$

$$\Leftrightarrow \frac{cos\left(\Delta\lambda\right) \cdot sin\delta \cdot sin\alpha}{sin\left(\Delta\lambda\right)} = cos\varphi_A \cdot cos\delta - cos\alpha \cdot sin\varphi_A \cdot sin\delta$$

$$\Leftrightarrow \quad \frac{cos\left(\Delta\lambda\right)}{sin\left(\Delta\lambda\right)} = \frac{cos\varphi_A \cdot cos\delta - cos\alpha \cdot sin\varphi_A \cdot sin\delta}{sin\delta \cdot sin\alpha}$$

$$\Leftrightarrow \quad tan\left(\Delta\lambda\right) = \frac{sin\delta \cdot sin\alpha}{cos\varphi_A \cdot cos\delta - cos\alpha \cdot sin\varphi_A \cdot sin\delta}$$

Inside the transformation cosine and sin rule as well as relation $sin^2x + cos^2x = 1$ are exploited. The result given in equation (A.3) can be solved for $\Delta\lambda$ by inverting the tangent since it only contains already known quantities δ, α and φ_A. It is independent of the already determined φ_B. However, replacing $cos\alpha$ by the formula applied before, it can be included easily:

$$tan\left(\Delta\lambda\right) = \frac{sin\delta \cdot sin\alpha}{cos\varphi_A \cdot cos\delta - cos\alpha \cdot sin\varphi_A \cdot sin\delta} \tag{A.4}$$

$$\Leftrightarrow tan\left(\Delta\lambda\right) = \frac{sin\delta \cdot sin\alpha}{cos\varphi_A \cdot cos\delta - \left(\frac{sin\varphi_B - sin\varphi_A \cdot cos\delta}{cos\varphi_A \cdot sin\delta}\right) \cdot sin\varphi_A \cdot sin\delta}$$

$$\Leftrightarrow tan\left(\Delta\lambda\right) = \frac{sin\delta \cdot sin\alpha}{\frac{cos^2\varphi_A \cdot cos\delta + sin^2\varphi_A \cdot cos\delta - sin\varphi_B \cdot sin\varphi_A}{cos\varphi_A}}$$

$$\Leftrightarrow tan\left(\Delta\lambda\right) = \frac{sin\delta \cdot sin\alpha \cdot cos\varphi_A}{cos\delta - sin\varphi_B \cdot sin\varphi_A}$$

$$\Leftrightarrow \quad \Delta\lambda = arctan2\left(sin\delta \cdot sin\alpha \cdot cos\varphi_A, cos\delta - sin\varphi_B \cdot sin\varphi_A\right).$$

The special version of the arc tangent function is used in equation (A.4) in order to be able to reproduce all possible locations of A and B on the earth's surface, for instance on northern or southern hemisphere or on the equator. The formula obtained is the same that can be found on [59], where no derivation is given.

Scientific Career

03/2008	Abitur at Hofenfels-Gymnasium Zweibrücken
10/2008 - 03/2012	Bachelor Studies of Mathematics
	with Computer Science as minor field of study
	at University of Kaiserslautern
	Degree: Bachelor of Science
04/2012 - 05/2014	Master Studies in Mathematics
	with Computer Science as minor field of study
	at University of Kaiserslautern
	Degree: Master of Science
since 07/2014	PhD studies at University of Kaiserslautern
07/2014 - 12/2017	PhD scholarship at department *Mathematical*
	Methods in Dynamics and Durability at Fraunhofer
	Institute for Industrial Mathematics (ITWM) Kaiserslautern
since 01/2018	Research associate at Fraunhofer Institute for
	Industrial Mathematics Kaiserslautern, department
	Mathematical Methods in Dynamics and Durability