A methodology for the extrapolation of trip chain data

Dissertation

an der Fakultät V

– Verkehrs- und Maschinensysteme –

der Technischen Universität

Berlin

vorgelegt von

Diplom-Wirtschaftsinformatiker

Sebastian Schneider

aus Trier

Oktober 2011

D 83

Promotionsausschuss:
Vorsitzender: Prof. Dr.-Ing. Markus Hecht
Berichter: Prof. Dr. Barbara Lenz
Berichter: Prof. Dr. Kai Nagel

Tag der mündlichen Prüfung: 18. November 2011

Abstract

Modeling and simulations are core elements for the evaluation of policy and investment projects. Particularly with the transport sector, determining the expected transport demand is a centrally factor. These arise due to changes to the location of individuals and objects within the area of study. Should this demand be established for a future period via prognosis, measures can be evaluated prior to their implementation.

Traditional approaches follow the macroscopic '4 step model', and ascertain demand in transport in aggregated form, for example via breaking them down into traffic zones and homogenous user-groups. Using microscopic simulations as a new branch of research however, each traveler is seen individually. They possess an important advantage – they are able to draw back complexity, found with modeling systems, to more simple rules, because complex phenomena at the macro-level are the result of the interaction found at the individual level.

The development of microscopic models generates many challenges. Efficient algorithms must be developed, as, according to the quantity and complexity of the simulated objects, the amount of calculatable individual cases will be great. In addition, advanced data requirements will need to be made, as for each individual an initial activity-plan is necessary – one which then can be followed and optimized during the simulation run. The difficulty therein lies in the creation of separate plans for each individual out of general input data.

As a solution, a method has been developed which applies representative, empirically-observed trip chain patterns to a chosen area of study. Trip-chains thus generated by the model exemplify themselves by the detail found within space at the level of individual buildings and down-to-the-minute time periods. Further, the trips generated remain connected and are thus suitable, unlike traditional origin-destination matrices, for the direct application within microscopic vehicle flow models.

Using Geographic Information Systems (GIS), the area of study is described and incorporates, according to the availability of data, a representation of the studied area via land-use data and elements such as the location of buildings, companies, and private households. The data concerning the behavioral patterns describe the pattern

of movement for individuals in the form of empirically observed trip chain protocols. Both data types can then be linked by the attributes that they have in common, expressed as sets of constraints.

The method's application will be demonstrated as an urban transport model observing commercial passenger transport within the Berlin region and the precise detail of the generated trip chains via comparison to established comparative data will be confirmed. It can be shown that the method is applicable and correct values for established key values of reference can be produced.

Zusammenfassung

Modelle und Simulationen sind ein zentrales Werkzeug für die Bewertung von politischen Maßnahmen und Investitionen. Speziell im Verkehrssektor stellt die Bestimmung der erwarteten Verkehrsnachfrage eine zentrale Fragestellung dar. Sie ergibt sich aus der Ortsveränderung von Individuen oder Verkehrselementen in einem Untersuchungsgebiet. Lässt sich diese Verkehrsnachfrage für eine bestimmte Zeit prognostizieren, können Maßnahmen bereits vor ihrer Implementierung bewertet werden.

Traditionelle Ansätze verfolgen dabei das makroskopische "4-Stufen Modell" und bestimmen die Verkehrsnachfrage in aggregierter Form, zum Beispiel durch Unterteilung in Verkehrszonen und homogene Nutzergruppen. Mikroskopische Simulationen als neuerer Forschungszweig betrachten hingegen jeden Nachfrager einzeln. Sie haben den wichtigen Vorteil, daß sich mit ihnen die Komplexität des zu modellierenden Systems auf einfache Regeln zurückführen lässt. Komplexe Phänomene auf Makroebene gehen dann auf die Wechselwirkungen zwischen den Einzelereignissen zurück.

Bei der Entwicklung mikroskopischer Modelle stellen sich mehrere Herausforderungen. Zum einen müssen effiziente Berechnungsverfahren verwendet werden, da je nach Zahl und Komplexität der simulierten Objekte die Zahl der zu berechnenden Einzelfälle sehr groß wird. Außerdem müssen unterschiedliche Datenanforderungen berücksichtigt werden, denn für jedes Individuum ist ein initialer Aktivitätenplan erforderlich, welcher dann im Laufe der Simulation verfolgt und optimiert wird. Eine Schwierigkeit dabei ist, aus den meist allgemein vorliegenden Eingangsdaten individuelle Pläne für jedes Individuum zu erstellen.

Zur Lösung wird eine Methodik entwickelt, welche empirisch beobachtete repräsentative Verhaltensdaten auf einen Untersuchungsraum überträgt. Die dabei erzeugten Wegeketten zeichnen sich durch eine feine räumliche und minutengenaue zeitliche Auflösung aus. Des Weiteren bleiben die erzeugten Fahrten durch alle Modellschritte hinweg miteinander verknüpft und eignen sich so, anders als herkömmliche Quelle-Ziel Matrizen, für den Einsatz in mikroskopischen Verkehrsflussmodellen.

Der Untersuchungsraum wird dazu in einem Geographischen Informationssystem (GIS) beschrieben und umfasst je nach Datenverfügbarkeit eine möglichst genaue Dar-

stellung des Raumes durch Angaben zur Flächennutzung sowie der Verortung von Objekten wie Gebäude, Firmen und Haushalte. Die Verhaltensdaten beschreiben die Bewegungsmuster der mobilen Individuen innerhalb der zuvor definierten räumlichen Umgebung und liegen in Form von Wegekettenprotokollen vor, die typischerweise in empirischen Studien oder durch fahrzeugseitig installierte elektronische Bordsysteme erhoben werden. Beide Datenarten werden dann durch ein auf deren Schnittmenge abgestimmtes Regelwerk verknüpft.

Die Funktion des Verfahrens wird für den Raum Berlin als urbanes Verkehrsmodell für den Personenwirtschaftsverkehr demonstriert und die Genauigkeit der erzeugten Wegeketten durch Vergleich zu etablierten Vergleichsdaten bestätigt. Es kann gezeigt werden, dass das Verfahren anwendbar ist und stimmige Werte für bewährte Kennzahlen erzeugt.

Contents

List of Figures

List of Tables

Listings

1 Introduction

1.1 Objective and relevance of the thesis

Annually, enormous sums of money are spent on transportation infrastructure projects around the world. Budgets in the hundreds of millions of dollars are common [34]. Assessment of both the financial potential and the risk of such projects heavily depends on the accuracy of traffic forecasts. This thesis contributes to the task by presenting a novel approach to transport demand generation with which to extend existing forecasting models. The approach is inspired by the fields of computer science and cartography and, as such, relies on computer algorithms, database systems and geographic information systems (GIS).

Traffic forecasts are employed in various fields. The need dates back to the 1950s [52], while the growing awareness of environmental issues further increased demand [33]. However, the forecast accuracy remains the key aspect. For example, Bangkok's Skytrain, an urban rail system costing US $2 billion (€1.5 billion, as of November, 2009), was designed to face passenger volumes 2.5 times higher than the actual traffic. Misleading traffic forecasts resulted in over proportioning: oversized station platforms and a large surplus of dispensable trains and cars that now queue in the system's maintenance areas [34].

For these reasons, transport modeling as an essential tool for forecasting require fine-tuning. Most of the research has, so far, distinguished between 'personal transport' and 'freight transport', as these two transport types make up the substantial share of all transport. However a third type, namely individuals traveling for commercial reasons and on business purposes, has been disregarded so far [61]. Growing evidence convinces however, that there is a necessity for considering it, for the following reasons:

1. First, since only about one third to a half of the commercial passenger transport trips are actually reported in written household surveys [19], a major share of relevant trips is not represented in passenger transport statistics. As a consequence, commercial passenger transport is only partly covered by existing personal transport models, which are typically based on passenger transport statistics.

2. Second, already today the share of commercial passenger transport seems to account for roughly twenty percent of the total urban traffic [50, 82, 101]. Hence, the fact that it is not yet considered within transport models is problematic.

3. Third, with regard to future developments, there are reasons to believe that its share, especially in the industrialized countries, is going to grow [102]. Evidence can be found in many articles hinting at future economic development. Thus THE ECONOMIST believes that Germany's service sector is currently underdeveloped and argues that the share of services sold domestically must rise [64]. The argument is that otherwise the gap that emerges as Germany cannot rely on export as much as it did in the past will not be able to be filled. A growing significance of the service industry would likely be accompanied by an increase in traffic of that industry, and only a share of the trips can be replaced by information and telecommunication technologies.

4. Fourth, recent model designs themselves make the consideration of commercial passenger transport necessary, especially in the light of the trend towards regional and spatially diverse model designs that provide results at the level of city districts and individual housing blocks. First studies indicated urban traffic to be the largest share of commercial passenger transport, thus regional models suffer the most if commercial passenger transport is not considered.

Besides the above reasons, the need to direct research towards taking commercial passenger transport into account has been repeatedly emphasized in scientific publications, among them by MENGE AND LENZ [61], STEINMEYER [88] and WERMUTH ET AL. [102], and it has further been acknowledged by official authorities such as the Berlin Senate [82], including the immediate need to develop models for commercial passenger transport [87].

The goal of the present thesis is to contribute to the solution of the problem by presenting a novel methodology that can be used for the synthetic generation of transport demand. This thesis' approach follows the principal idea of extrapolating empirically observed trip chain data through the reproduction of its main characteristics, while at the same time matching given geographic characteristics of the study area. For output, synthetically generated trip chain data is obtained for being fed into microscopic vehicle flow simulations, with the benefit that all trips in a trip chain are related and remain connected. As such, the methodology is designed to connect with simulations such as SUMO [29], MATSIM [55], and TRANSIMS [94] that see individuals as their

primary modeling unit. By converting the individual trip chain data to aggregated origin-destination (OD) matrices of cell zonings in lesser detail, the methodology can alternatively be integrated into traditional network equilibrium models for the generation of aggregate travel predictions. The methodology may also be used for transport types other than commercial passenger transport outlined in this thesis.

Furthermore, multi-agent simulations, in which each traveler is represented individually, often use the concept of systematic relaxation in order to implement feedback learning. A loop structure typically executes the same calculation many times, with the difference that the result of the previous run is fed into the current one until a solid state is reached [91]. To begin with, appropriate initial demand patterns are needed and for generating these, the present methodology is a logical tool. As such, the methodology contributes to the challenge formulated by Balmer et al. "to create individual demand patterns out of general input data" [7].

The principal design of the methodology is founded in the main observation that personal transport models typically cover trips that are made by individuals, while most models for commercial transport were designed to emulate flows of goods. As commercial passenger transport primarily deals with mobile individuals who only optionally carry goods, the methodology is based on that concept of personal transport modeling. A distinct characteristic of the present approach is that it relies on two main groups of input data: empirical trip chain data and spatial data.

1. The first group (trip chain data) is typically part of empirical mobility surveys, in which individuals are asked to report all their movements within a single day. Such surveys provide information on the type and the duration of the activities that were performed, as well as information on the trips between them, such as the trip's mode, duration and distance. Surveys of this kind are often constructed nation-wide on behalf of federal authorities or, on a smaller scale, on behalf of regional and/or private transport providers.

2. The second group of input data (spatial data) seeks to describe spatially the study area with characteristics that can also be found in the trip chain data. Benenson and Torrens point out that spatial data is accessible through national databases for most industrialized countries, but up to this time not much of its potential has been used for transport modelling, although modern methodologies such as airborne and satellite sensors provide a large set of information on urban areas at the resolution of separate buildings [10]. Despite these possibilities,

most current traffic models cut space down into aggregated traffic zones and then conclude how much traffic flows between the zones.

While an increase in spatial precision of the input data does not necessarily improve a model's output, there are several advantages with it. A major benefit is that with lesser degrees of abstraction, a model's comprehendability increases for others from outside the subject area. And further, finer breakdown leads to a greater variety of policy options that can be resolved and subsequently evaluated with the model [9].

The methodology is inspired by a background in computer science. On the implementation side, the approach employs agent-based simulation techniques and relies heavily on spatial database technology. The methodology utilizes constraints that can be directly expressed in Structured Query Language (SQL) statements for relational database systems.

1.2 Structure of the thesis

To begin with, the main characteristics of commercial passenger transport will be presented in chapter 2. A definition will be given along with a literature analysis of known characteristics, and, with this, the basis for the actual methodology design will be established. Chapter 3 will then summarize the principles of modeling and of transport modeling in particular. Chapter 2 and chapter 3 together form the basis of the feasibility analysis that is presented thereafter in chapter 4, bringing together aspects specific to commercial passenger transport and those pertaining to model theory.

The analysis pioneers a way to the actual methodology, which is developed in chapter 5 as a set of mathematical formulas and algorithms. Chapter 6 then presents the implementation into software, demonstrated by the example of Berlin, Germany. The results obtained are used afterwards for verification. First, it will be shown that the methodology can be successfully implemented and applied to a real world scenario, showing that its complexity is maintainable and runs in affordable time. The results further show that the output data consisting of individual trip chains is sound in that it corresponds to known aggregates of that region. At the end of chapter 6, a set of visual examples is given that highlight the key aspects of the generated data. Chapter 7 then assesses the advantages and disadvantages of the present methodology and investigates into the method's usability for scenario analysis and forecasting. The thesis concludes with a number of summary conclusions and a discussion of future prospects in chapter 8.

2 Analysis of commercial passenger transport

There are several approaches to transportation. A broad definition is provided by Pirath, who defines transportation as the movement of people, freight and information from one location to another [75]. Excluding the transportation of information leads to the more narrow concept that divides transportation into the two fields 'transportation of freight' and 'transportation of people', nowadays also referred to as commercial transportation and private transportation [103]. However, there are difficulties in clearly separating commercial and private transport, especially with regard to people traveling for commercial reasons. This particular type of transport can be referred to as commercial passenger transport, to which Steinmeyer dedicated her doctoral thesis in 2004. This chapter is oriented to her findings. In section 2.1, it first presents a definition that was suggested by her after analyzing several definitions from the scientific literature. Thereafter, the main characteristics are presented in section 2.2.

2.1 Definition of commercial passenger transport

Private transportation covers trips that meet private needs and that are typically made by the person having the need, and on the other hand, commercial transportation covers trips that are made for commercial activities, implying that the trip's intention is not the immediate satisfaction of the private needs of the person traveling. Wermuth et al. break down the latter category into the following sub-classes [102]:

1. **Freight transport** (or goods transport) refers to the transportation of freight. It includes trips made primarily for the purpose of transporting freight and is further split into:

    a) 'Trips on behalf of others' (or commercial freight transport), i. e. goods that are transported between the place of production and consumption, and

    b) 'Trips on own account' (or work transport), i. e. trips that transport a company's own goods on its own account, e. g. between production facilities.

2. **Commercial passenger transport** refers to traffic generated by people who travel for a business or official purpose. These trips can include the delivery of goods, tools, materials, equipment and the like as a secondary function, but not as the primary objective. Trips of the type 'commercial passenger transport' are caused by the non-physical nature of a service, which does not allow the product to be shipped as if it were a physical object [59]. Examples are delivery and installation services (if done together) or after-sales and repair services. Commercial passenger transport can be further broken down into service trips and business trips:

    a) Service trips (or service transport), which represent a combination of goods and passenger transport, where, besides the person actually providing the service, further materials such as tools or spare parts are carried. Examples are customer service providers and craftspeople.

    b) Business trips (or business and service transport) are those made for a business or an official purpose without the carriage of freight. Examples are trips for attending conferences, business meetings and the like. Steinmeyer further suggests classifying trips of this type according to their length in space and time [87]:

        i. Regional business trips refer to trips in which the reason for traveling is to perform a commercial task such as a meeting with customers within the range of 100 kilometers. In addition, the outward journey and return trip must occur within 24 hours.

        ii. Non-regional business trips refer to trips in which the reason for traveling is to perform a commercial task and which exceed the regional periphery of 100 kilometers and/or last more than 24 hours. Examples are customer meetings or visits to trade fairs.

3. **Transport service trips** (or passenger transport), which are made for the purpose of transporting people other than the person operating the vehicle. Examples are trips by empty taxis and buses. This is because, if not empty, the trip would be categorized by the purpose of the passenger making the trip.

For all categories, empty return trips, such as those back to the company's site or to a parking lot, fall under the same category as that of the preceding trip. The definition does not restrict commercial passenger transport to specific modes, nor is the geographical setting of importance. The main criteria are the person traveling

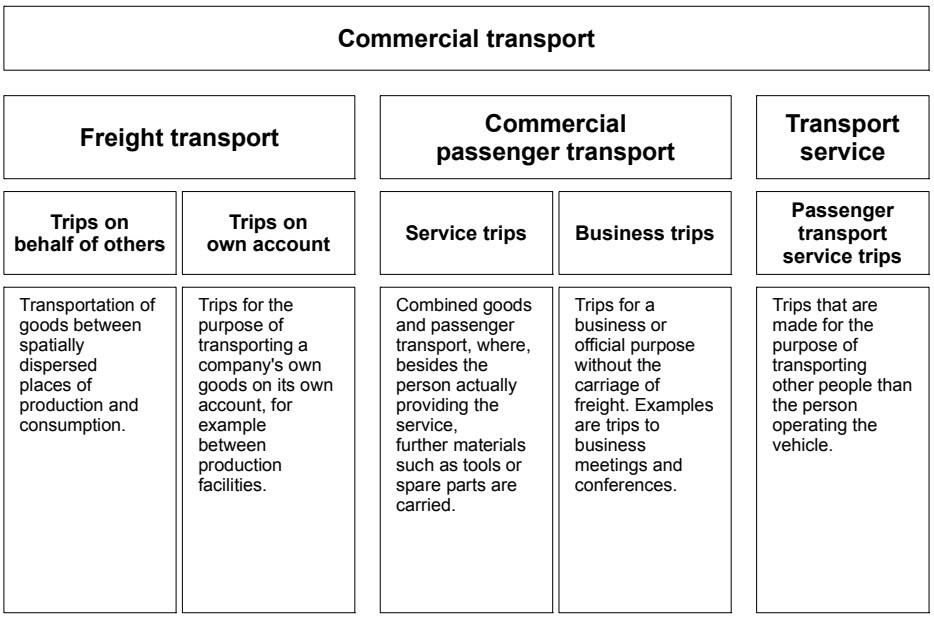


Figure 2.1: Commercial transport by Steinmeyer and Wermuth et al., own translation based on [87] and [102].

and the activity that causes the person to travel, while the transportation of freight is understood as a non-mandatory, secondary function [87]. Figure 2.1 shows a graphical representation first presented by Steinmeyer in German [87]. The figure shows how commercial transport is split into the categories presented before.

Her definition resulted from an analysis of German and international literature and borrows partly from the doctoral thesis of Schütte [81]. The empirical survey *Kraftfahrzeugverkehr in Deutschland* (KiD), which is of great importance to the present thesis, corresponds to the classification in figure 2.1 except for the introduction of an additional category. To account for the fact that definitions must be subjected to pragmatic considerations if they are to be used in empirical surveys [102], the category 'other commercial transport' is added for cases in which it cannot be distinguished clearly whether a trip should be categorized as part of freight transport or commercial passenger transport. Reported trips made for a combination of purposes or for a purpose that is not covered by the previous two categories belong in this category. This

applies for example, to trips made to maintain the operability of a vehicle such as trips to the garage or the gas station.

2.2 Characteristics of commercial passenger transport

Besides national statistical data, STEINMEYER analyzed several regional studies, in particular two empirical surveys for the cities of Hamburg and Dresden [87]. Since most of her data was part of regional studies, the findings are restricted to a regional context:

- The number of people that are affiliated with commercial passenger transport represent only a small percentage of the entire population, but show higher-than-average trip rates. Thus, they are extremely mobile.
- While commercial passenger transport is not restricted to particular modes, the major share is on motorized transport.
- Significant differences in terms of trip purpose, destination choice and means of transport can be identified depending on the economic sector of the businesses that generate trips for providing the service.
- There is a linear correlation between a firm's number of mobile employees and a firm's total number of employees, and the number of vehicles that a firm owns also correlates to it.
- Not only automobiles registered to commercial keepers, but also a considerable share of automobiles registered to private keepers produce commercial passenger transport trips, and a large share of commercial passenger transport originates from small companies such as roofers, painters and plumbers.
- While private households are usually not represented in freight traffic models, they are relevant for commercial passenger transport, since they can serve as possible destinations of commercial passenger transport trips.
- Urban commercial passenger transport occurs mostly in the near vicinity and a major share of it consists of short-distance trips.
- Commercial passenger transport is assumed to account for 19 % to 30 % of the total traffic in most studies. Hence, the impact of commercial passenger transport on urban traffic seems to be higher than the one of freight transport.

3 Principles of transport demand modeling

The previous chapter named the definition and the characteristics of commercial passenger transport. This chapter introduces modeling aspects in order to pave the road towards the development of a methodology that allows to model the corresponding transport demand. The two chapters will be connected by the feasibility analysis thereafter in chapter 4, outlining the principal considerations that lead to the actual methodology that is then developed in chapter 5.

The chapter starts by providing in section 3.1 a short overview on general aspects of modeling, showing that models became an important tool in most scientific disciplines and names some principal challenges during model development. Emphasize will be given to the importance of computer science in the light of current state-of-the-art modeling efforts. The chapter will then explore concepts specific to the field of transport modeling in section 3.2. The standard four-step model for transport modeling as well as the recent trend towards microscopic simulation systems will be outlined. Further, transport demand will be defined followed by a description of the activity-based travel theory, with respect to the importance of the two concepts for the present approach. The chapter ends by presenting a brief description of a simulation framework into which the present work can be embedded.

3.1 The use of models in science

3.1.1 Introduction to modeling

Models became popular in almost all fields of science and are one of the principal instruments of modern science. Their application ranges from the simulation of stars and galaxies, the simulation of nuclear reactions, the evolution of life to simulating the outbreak of wars [42]. Well-known examples are the Bohr model of the atom, the double helix model of DNA and the general equilibrium models of markets. A surprisingly large number of historic models use mechanics and are on display in museums today, such as the Phillips hydraulic model of the Keynesian economic theory, in which water

is used to represent the circulation of money [74]. By changing the rate of the flow of water through pipes, one can analyze effects such as an increase of interest.

Models can serve two fundamentally different functions [36]. First to represent a theory in a way that it interprets the laws and axioms of that theory. Second to represent a selected part of reality with the advantage of having the ability to run tests within the model even if a test in reality is not feasible, for example if tests are too costly, take too long or to run and test different scenarios against each other, enabling an investigation into the effects of changing parameters before applying them in reality. In 1973, STACHOWIAK presented a definition of modeling that is applicable to any scientific field [83]. According to his understanding, a model needs to incorporate three aspects:

1. **Emulation.** A model must always be a reproduction of something else, which in turn can be a model itself.

2. **Reduction.** A model addresses relevant aspects only. All other aspects are suppressed and do not become part of the model.

3. **Pragmatism.** A model is developed to fulfill a specific need, perhaps with just a limited period of time in mind. Its design should be oriented for addressing this need.

Two main aspects can be named that are crucial for model development. The overall goal is to express as clearly as possible the way in which one believes reality to operate [37] while another goal is to transfer the complicated real world to a simpler scale so as to make the emulation possible. The difficulty is to find a convenient ratio that achieves both goals at the same time.

3.1.2 Computerized modeling

Models do not necessarily need to be very complex. Instead, a simple regression equation between an independent and a dependent variable can already represent a model, like for example the Cobb-Douglas production function[1] $Y = AL^{\alpha}K^{\beta}$ that interprets how manufactured products relate to production input factors [23, 37].

While models can be a mathematical formula or a physical object, the computer offers a wide range of possibilities to the modeler, making those implemented on a

1 with Y = total production, L = labor input, K = capital input, A = total factor productivity, α = output elasticity of labor, and β = output elasticity of capital

computer a highly effective methodological tool [42]. Therefore, it is not surprising that most models nowadays rely on computer technology in one way or another. As a consequence, the overlap of computer science and modeling (and traffic modeling in particular) is manifold. It ranges from research on efficient algorithms over processing large data matrices or n-dimensional data cubes using appropriate data structures onto designing, selecting and assembling hardware systems that provide sufficient computing power.

Many concepts for computational modeling were originally developed in other scientific fields than the field of application and are borrowed from computer science, physics, mathematics and economy. Some of the most important concepts that are borrowed from computer science are automata theory, multi-agent systems, artificial intelligence, and object oriented programming [10]. For transport models in particular, graph theory, databases and geographic information systems (GIS) are heavily used [80].

There is a general agreement that computational performance is a major challenge for many model designs, in particular for multi-agent simulations in the transport sector, because lacking performance is seen as a major impediment to the acceptability of such computationally intensive models [9]. For these, overnight computation times on standard hardware are considered critical for such transport models to move from research labs to regular professional application [2]. Nagel and Marchal highlight the interaction between computer science and traffic research by pointing out that the development of a model and its implementation are not independent processes. Already during the model design phase, one must consider the computational techniques and the actual model logic together. This is particularly significant for the development of models in a sustainable and modular way, because once a particular implementation has been selected, a certain amount of lock-in toward certain models can be expected until someone else takes it on to start a competing, different implementation [66].

A beneficial side effect of computational modeling is that it forces the modeler to be precise, because unlike models expressed in natural language, valid computer code is always exactly specified if it is to run [37].

3.2 Models in transport research

In the specific field of transport research, models are deployed to forecast traffic and to estimate the effects of various policies. Bobinger distinguishes between two groups of traffic models [14]. Models of the first provide forecasts based on economic figures. Their main purpose is trend forecasting for large regions on macro-economic level, for example based on a region's gross domestic product (GDP). The second group stresses the spatial distribution of traffic and analyzes traffic flows between regions. Space is often cut into disjunct zones for this purpose, with each zone described by socio-economic characteristics such as the number of citizens, the number of work places, the sales area in local stores and so on. Zone-to-zone travel demand is then predicted for a given time period by executing several successive sub-models that account for the characteristics of the zones as well as for the distance, time or cost required to travel between them [9].

3.2.1 The standard four-step model

For models belonging to the latter group, the standard four-step model (FSM), sometimes also referred to as the urban transportation planning (UTP) procedure, was established (see for example [56, 26, 62]). The standard procedure, which, as the name already suggests, consists of four steps, provides a general framework for modeling transport. Due to its basic nature, the four steps can be identified in most transport models. The four steps are:

1. **Trip generation** determines the number of outgoing and incoming trips for each distinguished geographic area called zones. Zones are obtained by dividing space into multiple areas, ideally by considering traffic-specific aspects.

2. **Trip distribution** then calculates how these trips distribute by assigning destination zones to the outgoing trips for every zone that has outgoing trips. The outcome typically is stored in so-called origin-destination-matrices (OD-matrices) and will be discussed in more detail at the end of this section.

3. **Mode choice** determines which mode of transportation is chosen for each trip, e. g. whether the trip was taken by rail, by car, by truck or by bicycle.

4. **Route assignment** finally converts trips into routes. Here, usually a user equilibrium is sought that ensures that all paths that are used for a given origin-

destination combination have the same travel time and no other, faster travel path exists [99].

The four steps and the data that is obtained after each step is outlined in figure 3.1. Recent models tend not to strictly follow the individual steps in sequence, but allow some kind of interaction between them, paying respect to the idea that the steps should influence each other. Furthermore, trip generation, trip distribution and mode choice are sometimes carried out separately for different user classes (or homogeneous groups), with every user class representing individuals with similar travel behavior.

Although not identified as a separate step, it should be added that the collection and preparation of data is a prerequisite and can make up a substantial share of the work. This involves evaluating national statistics, but also employing designated empirical surveys, e. g. time-budget surveys or trip attraction surveys. If real-life data is missing, it is necessary to generate synthetic data that meets known aggregate statistics [9]. Fortunately, the amount of available data increased steadily during the past years. Nonetheless, this effect will not completely eliminate the need for generating synthetic data, because firstly, some data is subject to privacy requirements. Although existent, this data is then not available for research [10]. Secondly, for some data the extraction from reality by empirical means is impossible due to the kind of data that is needed, such as future data. And finally, the synthetical generation of data can be required due to technical difficulties that do not allow to obtain the needed data by other means, for example to survey personal dayplans of all individuals of a whole region during the same day.

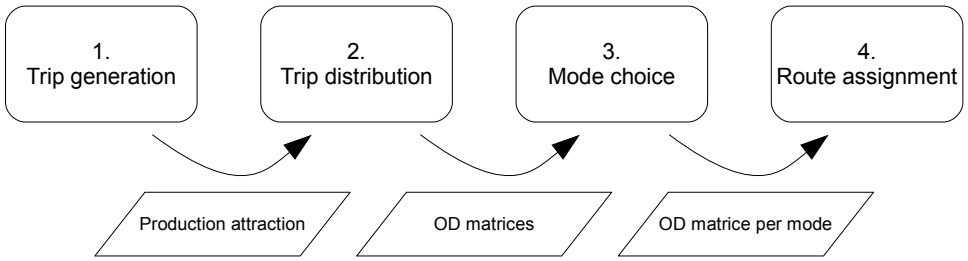


Figure 3.1: The four step process for transport modeling.

3.2.2 From macroscopic models towards microscopic simulation

Models in which individuals are aggregated into user classes and space is represented as aggregated zones are often referred to as macroscopic or aggregated models. The improved availability of data accompanied by the substantial increase in computer power in recent years has shifted modeling towards a more disaggregated level. Existing models have been refined by distinguishing between a greater number of user classes and by increasing the number of zones. However, Axhausen observes three main problems that are inherited in the principle design of such aggregated models [2]. Firstly, the approach does not allow proper emulation of regional and short term effects that occur within the transport network such as traffic jams, environmental pollution, traffic for finding parking space or waiting lines for public transportation. These generate substantial costs and should be taken into account by models. Secondly, modern requirements to transport modeling require individuals to be grouped according to highly differentiated classifications schemes. And thirdly, aggregated approaches face general problems if it is to emulate the entire decision process that individuals maintain for their daily planning. In response, a different field of research was established which focuses at individuals. In microscopic simulation, real world processes are seen as the result of individual objects interacting with each other [10], often at the scale of individual households or persons. Total results are obtained by aggregating the individual results that are based on individual-level characteristics.

A clear borderline between macroscopic and microscopic approaches is hard to draw. On the one hand, this is because the term microscopic simulation has been interpreted quite diversely in transport research [91] and on the other hand, many kinds of intermediate modeling approaches exist. One point is that microscopic simulation models generally focus much more on small units, going as far as resolving all elements of the transport system such as roads, crossings, vehicles, and travelers [12]. Opposed to this, macroscopy refers to a more global view on reality and emphasizes the flows of goods and travelers opposed to individual entities.

There are several advantages to micro simulation. One is that they imply the possibility to implement any logic without being restricted to mathematical relationships in the form of formulas, thus offering a greater variety of options to the modeler. The approach also allows to incorporate data at household or individual levels without information loss that may occur if input data have to be aggregated first before being fed into a model (however, the opposite is more common and data have to be synthetically disaggregated). Furthermore, micro simulation allows to aggregate the results into dif-

ferent aggregation levels, be it by time, by space or by mode type, and to output them either as aggregated statistics or as individual event listings, hence offering a higher range of output options [6]. The precision in resolution further makes such tools more sensible and more suitable for a wider range of policy options. And Ben-Akiva et al. point out the advantage that these tools allow to capture the impacts of traffic control devices including intelligent transportation systems (ITS) technologies, enabled by a detailed representation of the underlying traffic infrastructure [9]. Finally, these tools are more intuitive and therefore allow easier data interpretation, which makes them more suitable for explaining the results to those less familiar with the subject.

The microscopic simulation brings a number of advantages; nonetheless, some aspects of the reality make this type of simulation inapplicable in some cases. Microscopic models are typically bottom-up models where a lot of interaction occurs between small entities on the micro-level and then leads to aggregated effects on the macro-level [72]. Nevertheless, some processes work top-down, e. g. the designation of land-uses by governmental authorities or construction of highways [10]. Moreover, the bottom-up approach is more likely to be helpful for people trained in physics, engineering or computer science, whereas it may appear counterproductive for people trained in economics since they tend to take a top-down approach. [66].

Another disadvantage to micro simulation emerges when it comes to the practical application. At this time, macroscopic models have reached a more mature state than their microscopic counterparts, as they have seen continuous enhancement through constant development and practical use over the last thirty years that now is available as standard software implementations. On the contrary, microscopic simulation is a comparatively new field and at this time mainly used for scientific interests rather than real world applications [12]. Self-implementations must be developed in most cases, although modeling frameworks are available that intend to keep developing efforts low. Yet the extra effort did not prevent simulation-based demand models to become more widely used in recent years [18]. Ben-Akiva et al. state that “simulation has proven to be a flexible, versatile, and comprehensive tool for demand modeling” [9].

3.2.3 Formalizing transport demand

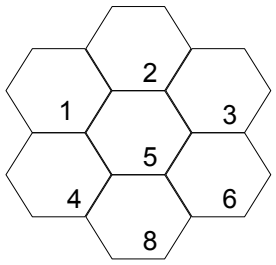
For both macroscopic models and microscopic simulations, a substantial milestone is obtained after trip generation and trip distribution, which is the demand for transport. After mode choice, this represents the input for the route assignment step which then ultimately results in flows. In economic model theory, demand is precisely defined

and commonly understood as the demand for produced goods for satisfying needs. SCHILLER observes that despite its significance, no precise definition exists for the field of transportation [80]. In consequence, traffic demand is understood here as the number of movements that are made during a given period of time. In the most generalized form, this includes all movements of people, vehicles, information or freight. It can be expressed in various aggregational forms, for example the total traffic volume of a given period or that of a specific individual in that period, or the traffic volume between two locations in space.

The traditional way to express transport demand is with origin-destination-matrices (or OD matrices). Such matrices specify the number of trips between zones for every zone pair as shown in the example in figure 3.2(a). One disadvantage of representing the demand for trips in form of zone-based matrices is that a matrix only represents one period in time. For example, a matrix can contain all trips between zones within one day. If a higher resolution is needed to resemble the dynamics of traffic within these 24 hours, several matrices can be created for several times of a day. This, however, disconnects trips that actually relate to each other. For instance, congestion might cause the succeeding trips of a vehicle to be delayed or the driver to re-schedule his dayplan giving up some trips at all, which cannot be resembled by conventional OD matrices. Further, microscopic simulation models that compute route assignment for individual travelers in effect require individual dayplans as input. If they are to be fed with OD matrices, the matrices must first be decomposed into individual records. Here, it is more efficient and conceptionally simpler not to use aggregated OD-matrices as input, but to start from individual trips [97].

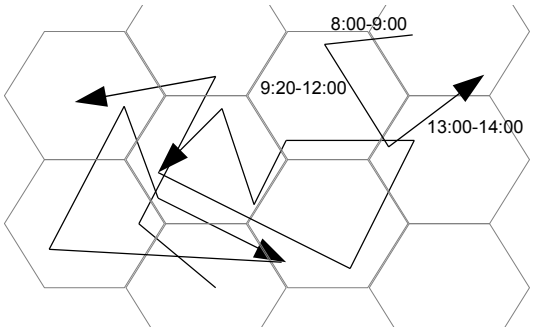
Trip chains as illustrated in figure 3.2(b) resolve these downsides. A trip chain contains the trips of an individual in sorted order, in consequence preserves the relations between them, e. g. outward journey and return trip, and can also be used directly for microscopic route traffic assignment. The transport demand of a region can be represented by storing all trip chains of that region. Trip chains are downward compatible, as merging all trips over a given period of time results in the conventional OD-matrix of that time. And by expressing transport demand through individual trip chains, the conversion that is necessary for feeding the data to microscopic simulation models is avoided.

Due to the complexity of modern transport models that consist of various sub models and modules for data preparation and visualization, many research groups do not have the resources to work on all aspects and hence concentrate their research on specific



$$\begin{matrix} v_{11} & v_{12} & \cdots & v_{1n} \\ v_{21} & v_{22} & \cdots & v_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ v_{n1} & v_{n2} & \cdots & v_{nn} \end{matrix}$$

(a) Transport demand represented as an origin-destination-matrix (OD matrix) for n cells. The values v_{ij} represent the traffic volume (number of trips) from the cell i to cell j.



(b) Transport demand represented as individual trip chains.

Figure 3.2: Alternatives for expressing transport demand.

parts [2]. In order to allow for collaboration between research groups, the need to come up with universal data exchange formats arises, for which Axhausen suggests the use of the XML file format following the example of the ddi initiative [25] for the social sciences or RailML [93], an initiative for the railroad sector in which among others, the German Aerospace Center (DLR) participates [2]. While likewise no standardized file format has been developed yet for the transport modeling community, the XML Document Type Definitions (DTD) provided by MATSIM, an open source toolkit for multi-agent transport simulations, can be used for the computerized representation of transport demand until a common standard is approved. The files are accessible in the current version at http://matsim.org/files/dtd/plans_v4.dtd.

3.2.4 Activity-based travel theory

The significance of trip chaining and the interest in the activities that trigger trips has steered research towards activity-based demand generation (ABDG). ABDG is based on the fundamental principle that travel demand is derived from activity demand [47]. Trips are seen as the result of activities that take place at different locations and the need for transportation then originates from the fact that two consecutive activities at different locations must be connected by a trip [97]. The concept implies that travel decisions are part of a broader activity scheduling decision process. As a consequence, understanding travel demand requires understanding the activities that trigger trips [9], and calls for modeling the underlying human decision process [57].

Critics of the activity approach question whether this level of model complexity, i. e. trying to recreate the complexity of human behavior, really is necessary for the need of predicting travel. However, the need to develop transport models that are more close to reality than is the case for the standard four-step model (FSM) is emphasized by McNally et al. who justify having to deal with the complexity, because "at this point, it is only possible to conclude that the current level of abstraction evident in the FSM is clearly insufficient, and that some enhancement, and probably a significant enhancement of the abstraction, is required" [57]. The two standpoints suggest that the best choice might be found in between: a system that is oriented more close to reality than the four-step model without the aim to resemble the human decision process.

Bowman and Ben-Akiva observe two central ideas in activity-based travel theory [18]. First, the demand for traffic is derived by the activities of people and second, individuals are further constrained in time and space. Considering the limited speed at which a person can move in combination with other activities that the person performs during the day, the space in reach during periods of travel is constrained. Hägerstrand first visualized the connection between space and time by so-called space-time-prisms [41]. Two examples are illustrated in figure 3.3. An important example is the natural need that under most circumstances, people return to a home base at the end of the day for rest and by that they return to a fixed place. Other activities that are performed during the day hence must be located within reach of the home base during the time available for travel [38]. The concept of activities can be passed on to zone-based traffic models by adding the attractiveness for various activities to a zone's description in order to influence trip generation and trip distribution. The concept can also be broader integrated by letting the activities be the main focus of observation, and then generating the resulting trips posterior. Specific activities such

as home, work, shop, leisure are assigned to individual users and then activity plans are developed that define when, where and in which order the activities are performed.

The work of Damm [24], Kitamura [49] and others provide reviews of the scientific literature on activity-based travel theory. By now, a considerable number of implementations of the concept are known. An examination of recent activity-based and non-activity based models that are currently in use or under development in the United States has been given by Rossi et al., stressing on differences between the two model types [79] and an overview on the historical development of activity-based modeling is given in [16]. In terms of European efforts, the Dutch model system ALBATROSS of the Technical University Eindhoven and MATSIM-T, developed in a Swiss/German joined effort by the Swiss Federal Institute of Technology Zurich and the Technical University of Berlin, can be named. An analysis of the two model systems has been given by Axhausen [2].

3.2.5 Static traffic assignment, dynamic traffic assignment and iterative demand optimization

Strippgen [91] and also Raney [77] explain the concepts of static traffic assignment, dynamic traffic assignment and iterative demand optimization and how these concepts relate. A brief summary will be given in the following in order to depict a framework in which the methodology that will be developed later on can be applied.

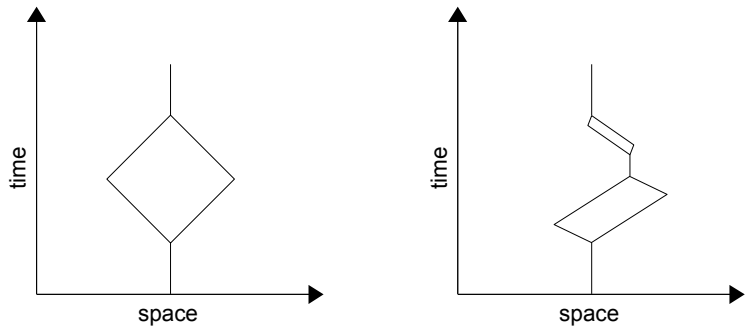


Figure 3.3: Hägerstrand space-time-prisms by [41] and [44]. A space-time prism is the set of all points that can be reached by an individual by a given maximum speed if both the starting point and the ending point are known.

Static traffic assignment determines the link volumes of a given network that result from a given transport demand during a given period of time. The network, which defines the supply side, consists of a set of links that are characterized by various attributes such as speed and capacity. The demand side typically is provided by an origin-destination (OD) matrix as described in section 3.2.3. Several approaches exist for the assignment ranging from incremental procedures to general equilibrium methods.

With dynamic traffic assignment (DTA), traffic flow models are used that describe the movements of vehicles in the road network in order to allow a more realistic modeling of the temporal dynamics of traffic flow. DTA models, opposed to static traffic assignment models, require activity plans directly to be fed into the traffic flow models for using the timing information that they contain. A significant drawback is that no stringent mathematical theory exists for DTA similar to the equilibrium solution of the standard four-step model, which is why multi-agent simulations can be seen as the preferred choice for computing traffic assignment. Another argument is that the multi-agent simulation approach well accompanies activity-based demand generation, as the first requires the individual plans that the latter provides.

With simulation-based traffic assignment, the concept of systematic relaxation is often used in order to implement feedback learning. A loop structure typically executes the same calculation many times, with the difference that the results of the previous run is fed into the current one until a solid state is reached. This is accomplished by executing the following steps:

1. Take an initial transport demand as input. For multi-agent simulation systems, this would be a set of dayplans that provides every agent of the simulation with one dayplan that the agent is going to follow.

2. Simulate the agent dayplans simultaneously.

3. Adapt dayplans for some or all agents by either choosing a different one or by keeping it but changing the route within the network.

4. Go on with step number two.

For micro-simulation models of this kind to work, Balmer et al. point out the challenge to create individual demand patterns out of general input data for every individual participating in the simulation [7], to which this thesis contributes the methodology that is developed in chapter 5.

4 Feasibility analysis

The previous chapter names basic principles on the theory of travel demand modeling. This chapter will derive requirements that will lead us to a specific model system. It is divided into four sections: section 4.1 will present general model theoretic requirements with an assessment of their practicability for commercial passenger transport; sections 4.2 and 4.3 will be dedicated to data sources since the availability of input data should be taken into account; section 4.4 states the conclusion leading to a specific model system, which, being originally designed to model transport demand for the service industry, can be applied for other transport types as well. The model itself will be introduced in detail in chapter 5.

4.1 Model theoretic requirements

Both private transport and commercial passenger transport have individuals as a principal unit, whereas freight transport models commonly deal with the flows of goods [59]. Therefore, the strategy for modeling commercial passenger transport should presumably be rooted in the theory of private transport modeling, focusing on the activity-based transport demand modeling approach.

To develop activity-based transport demand models, the set of requirements that is displayed in table 4.1 has been established by Bowman [15]. To begin with, he demands that a model be behaviorally and mathematically sound, since otherwise no basis would exist that allows us to rely on the results. He further requires a model to provide a resolution that is accurate enough "to capture behavior that affects the aggregate phenomena of interest" (including explanatory factors for choice). Third, it must be feasible to actually implement the model from a practical point of view. Hence, the necessary data must be available along with separate data which can be used for model validation and, in addition, one must be able to generate the input data synthetically for forecasting. Further, the model's logic and complexity should be computable and maintainable and the calculations should be performed in reasonable

time at reasonable costs. Lastly, he requests that the model produces valid results, that is, "the model must prove itself in validation" [15].

To meet all the requirements, ideal conditions are needed. Especially the choice component requires a solid conceptional and empirical understanding of the processes that are to be modeled. With regard to the service industry, this presumes to well understand the human decision process in that industry in order to develop a synthesized decision process that emulates choice appropriately.

Assuming that any commercial activity is ultimately subject to profit maximization (which can be precisely expressed through mathematical formulas), it can be argued that commercial decisions follow rules that are generally well predictable. Also commercial activities of the kind that is not tied to immediate monetary income could be considered by translation into monetary figures. On the other hand, there are aspects that must be set against the potential benefits of modeling the decision process. First, there is general agreement that in recent years research has been focused only on understanding the decision process of individuals for personal transport. In contrast, the behavior analysis for the commercial sector is less sophisticated and until now, many aspects of the decision process remain unknown [61]. Second, despite the principal simplicity connected with the clear goal of income generation, decision processes in the commercial world seem to be more complex than those of individuals in a private household context, when examined from a microscopic point of view (think of the variety in business models, product pricing strategies, or advertising campaigns). In this respect, Gressel and Mundutéguy analyzed the working conditions of mobile

Table 4.1: Requirements of activity-based travel demand forecasting models

1.	Theoretically sound for accurate results
	a) behaviorally
	b) mathematically
2.	Activity schedule resolution for policy sensitive information
	b) universal alternative set
	b) explanatory factors
3.	Practical resource requirements for implementation
	a) data for estimation, validation and model inputs
	b) maintainable logic (software)
	c) affordable computation (hardware)
	d) usable operator procedures
4.	Valid results

Source: Bowman [15]

workers in a number of economic sectors and highlight the discrepancy between a mobile worker's own personal interests and meeting the employer's interests while having to face the expectations of other, third-party actors encountered during their commercial activity [40]. The findings show a surprising level of complexity by unveiling an extensive system that makes it difficult to identify simple, universal and reproducible patterns. Third, critics of the decision modeling approach generally question whether it is really helpful to rebuild the human decision process for the institutional goal of travel forecasting, pointing to its complexity and multi-faceted nature [57].

A compromise can be reached with regard to the fact that large datasets of trip chains for the service sector along with information on the activities between the trips have been surveyed. Thus, in principle, activity schedules are available, but information on the decision process under which these schedules were planned is not. Hence, although this does not allow for a reproduction of the human decision process, the observed behavior can instead be taken for granted, believing that the individuals behaved as observed with a reason in mind. By taking trip chains as empirical input, the actual reason for action is not further questioned, as it seemed appropriate to the decision maker at that point, be it because of minimal costs, shortest route, time constraints, priority constraints, personal preference, service level agreements, chance or other reasons, which implies a generally lower risk of misinterpreting and, consequently, incorrectly recreating the underlying decisions. This also allows imperfect decisions to be part of the model, such as decisions of an arbitrary kind or wrong decisions based on false or incomplete information. As Hertkorn puts it, the observed behavior took place in reality and thus reflects how reality apparently operated at the time of the observation [44].

The prospect of integrating commercial passenger transport trips into models by utilizing empirical trip chain data suggests further investigation of a methodology for this purpose. Hence, various data sources are examined next, starting with empirical ones. With respect to the importance given to trip chain data, section 4.2 focuses on sources for trip chain data and ends with a characterization of the two data types 'population data' and 'traffic volume data'. Due to the fact that the latter two are common input data for other transport models, the corresponding sections are kept small. Section 4.3 is then dedicated to spatial data with the intention to develop a model system drawing heavily from cartographic data in order to compensate the default of data on commercial passenger transport.

4.2 Sources for empirical input data

When developing transport models, one typically faces a large variety of input data, which can differ in quality, spatial resolution, purpose, etc. [7]. The common approach is to combine these various data types in a suitable way, so as each of them contributes to the accuracy of the output results. Three kinds of empirical input data are presented in the following. To begin with, two sources for trip chain data are named. One example of traffic volume data and population data for each is given afterwards.

The intention of this section is to provide an overview of the empirical data that is available in order to pave the road towards the methodology that is to be developed subsequently. Some of the data sources that are introduced here will be used in the demonstration example in chapter 6. Of great importance for the development of the methodology in chapter 5 is the observation that the data sources share a common set of attributes with which they are linked to each other. So is the economic sector and the firm size of the vehicle holder part of the trip chain survey *Kraftfahrzeugverkehr in Deutschland* (KiD). The same attributes are also given in the traffic volume data and also appear in the population data. For now, an overview of the principal characteristics of the data sources will be given. How they are linked in the light of the methodology that is developed subsequently will be explained during model development in chapter 5 and will further be part of the demonstration example in chapter 6.

4.2.1 Trip chain data

A surprisingly high number of sources for trip chain data can be found, many of which are either not well documented or restricted. Consequently, this section does not aim to provide an exhaustive list, but will give two examples. The first is a major transport survey in which (1) the German service industry participated and (2) that is available to the public:

Kraftfahrzeugverkehr in Deutschland (KiD)

The empirical scientific survey *Kraftfahrzeugverkehr in Deutschland* (KiD) is a representative survey on the commercial motorized transport in Germany and it was carried out on behalf of the Federal Ministry of Transport, Building and Urban Development (*Bundesministerium für Verkehr, Bau- und Wohnungswesen*, BMVBS) in 2002/2003. KiD is at present the most comprehensive and meaningful data source on commercial

passenger transport [59]. While the conceptional and methodological design of the study is layed out in detail in [100], a short summary will be given here.

The main focus of the study is on motor vehicles with up to 3.5 tonnes of payload. Other vehicle types were considered in smaller numbers so as to analyze the full spectrum of commercial transport and be able to merge the results with other major traffic surveys [28]. Vehicle operators were requested to track all of their vehicle movements for a period of one day, which resulted in an ordered set of trips for each vehicle. Trips are characterized by distance, destination type, purpose, time, number of passengers, freight, etc. Further, geographic coordinates are on record for each waypoint. Their precision varies and can range from the center of a postal code area to pointing to the center of a road segment in accordance to data privacy requirements. The random sample for KiD was drawn from the Central Vehicle Register (*Zentrales Fahrzeugregister*, ZFZR) by the German Federal Motor Transport Authority (*Kraftfahrt-Bundesamt*, KBA), and the survey provides a detailed list of attributes from the register, ranging as far as the cubic capacity, the gross vehicle weight, the rated power, the maximum speed, the number of passenger seats, whether the vehicle is equipped with a hitch and so forth for every surveyed vehicle. The ZFZR register collects vehicle and owner data sent by the local registration authorities and insurers. At this time, approximately 50.2 million motor vehicles and 5.9 million trailers are registered [51]. The register is exhaustive and up to date with the limitation that it does not include foreign vehicles. In consequence, KiD reflects only those vehicles that are actually registered in Germany and incoming traffic from foreign countries as well as going-through traffic is not accounted for. Due to practical considerations, the vehicles were not entirely surveyed on the same day, but were surveyed either on a Tuesday, a Wednesday, or a Thursday in all cases. If the owner of a vehicle is a firm, it is further characterized by its business sector as well as the firm's size.

For the demonstration example in chapter 6, the survey will supply trip chain data to a simulation. A remake of KiD is expected for 2011 and is going to have a similar design as the previous survey of 2002/2003 (see www.kid2010.de). Accordingly, what is laid out in this work can be directly applied to the new data once it is available.

Trip chain data from logging units

Hu et al. point out that most transport studies focus on the theories and methodologies of travel behavior and do not investigate the possibilities to access novel forms of travel data, although such data nowadays often accumulates in large volumes [45].

Therefore, a different and less conventional source for trip chain data is exemplified in the following: Trip chain data is also often generated by firms for their fleet management or as a requirement for tax compliance, and many standardized industry solutions exist for this purpose. One is 'TomTom WORK' (www.tomtomwork.com), a system consisting of on-board units installed in each vehicle and a centralized fleet management software that informs on the operations of the connected vehicles. Logbooks featuring driving times, trajectories and a trip's purpose are retained. While information on the first two is obtained automatically through the on-board unit, information on the third is given by the driver. The software provides reporting capabilities in form of human readable text files and comma-separated values that allow a computerized analysis. Also noteworthy is that the system is designed to provide users with comfortable interfaces to enter the relevant information (e. g. a remote control with a separate button for each trip purpose), which suggests that the collected data is sound.

Due to competition and privacy reasons, firms typically will not make their internal information freely available, which makes the alternative more useful for joint projects in which firms and researchers collaborate. The advantage of such logging systems is that they combine high surveying precision in space and time while providing data over periods from one day to several months. In addition, they are accompanied by relatively low costs, as the survey does not originally take place on behalf of the research body and fulfills other primary interests.

Despite the benefits in comparison with conventional traffic surveys (namely reduced costs and high case numbers), data of this kind does not receive adequate attention [45]. The given example is just one of several tracking systems available on the market and shows that trip chain data is collected by firms every day.

4.2.2 Traffic volume data

The traffic volume refers to the actual number of trips that are made during a given period of time and it can be obtained in its most general form by counting the vehicles on the road network, be it manually or automatically. For this purpose, most countries have traffic counting stations installed on many important road segments from highways to urban areas.

Many counting stations not only tell the total number of vehicles that pass on a road segment, but also distinguish between several vehicle types. For example, the number of motorcycles, automobiles, trucks, semitrailers and others can be obtained. However, when it comes to detecting the traffic volume with respect to specific trip purposes

(such as the number of vehicles that operate for the delivery of a service), empirical surveys are the only method for data collection, because the range of different vehicle types used for a given purpose and the conformism between automobiles operating for private versus commercial purposes does not allow to determine a vehicle's trip purpose with certainty [87].

With regard to commercial passenger transport, only few statistics exist that give insight into the actual traffic volume that is generated (see chapter 2), which is why a survey specifically dedicated to commercial passenger transport was carried out in 2005/2006 in a joint effort by a number of research units, among them IVT (*Institut für angewandte Verkehrs- und Tourismusforschung e. V.*) and the German Aerospace Center (DLR). Detailed information on the project *Dienstleistungsverkehr in industriellen Wertschöpfungsprozessen* can be found in [43]. The representative study considered a sample of 2 313 companies in Germany. In order to account for Germany's economic structure with a relatively high number of small companies, the random sample was drawn favoring larger firms and specific economic sectors [61]. Unlike KiD, the survey was not designed to collect trip chain data, but to estimate the amount of traffic generated by firms [43]. In section 6.4, it is shown how mobility ratios are derived from it that specify the amount of traffic that is generated by a firm according to its economic sector and its firm size, where firm size is measured in number of employees. Besides, several future trends were indicated by the study, including a higher market penetration of information and communication systems as well as combined navigation and disposition systems like 'TomTom', which was introduced earlier in section 4.2.1.

4.2.3 Population data

Another data type that is typically required by activity based transport models is population data. A model system for commercial passenger transport demand generation requires population data for both firms and private households. While the former are the principal producer of commercial traffic, the latter are relevant because many commercial trips lead to private households. Despite the fact that, in principle, population data can be obtained directly from the real world, it is synthetically generated in most cases for two reasons: The first is for forecasting, i. e. to have a methodology that converts current data into future data according to expected future developments. The second is to account for missing data, for example if higher aggregated data must be disaggregated. This is often the case if data on individuals is protected by privacy requirements or if the collection of individual data is too costly. As a requirement

for the disaggregation process, the resulting data should feature statistical properties similar to the aggregated data.

Up to now, no general methodology for the synthetical generation of population files applicable to any country, area or setting has been developed and neither does standard software exist that guides through the process. One reason for this is that the design is very much determined by the quantity and quality of the available data, which tend to vary significantly and must be analyzed individually for each setup. In addition, different model designs often require different approaches to the synthetic generation. The literature reports on a variety of systems. See for example Beckman et al. for the creation of synthetic populations using American census data [8], the doctoral thesis of Hertkorn for an example of the city of Cologne, Germany [44] or the general description on the generation of synthetic population data by Bowman [17].

The approaches are often based on iterative proportional fitting (IPF) [27], which can be best described for the two-dimensional case with two variables. The methodology adjusts a table of data cells such that they add up to given totals for both the columns and the rows of the table. The unadjusted data cells can be referred to as the seed cells, and the given totals can be referred to as the marginal totals. Formally expressed, this is the matrix

$$\begin{array}{cccc|c} c_{11} & c_{12} & \cdots & c_{1j} & t_{1\cdot} \\ c_{21} & c_{22} & \cdots & c_{1j} & t_{2\cdot} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ c_{i1} & c_{i2} & \cdots & c_{ij} & t_{i\cdot} \\ \hline t_{\cdot 1} & t_{\cdot 2} & \cdots & t_{\cdot j} & t \end{array} \tag{4.1}$$

, where c_{ij} are the seed cells, $t_{i\cdot}$ are the margin totals of the first variable and $t_{\cdot j}$ are the margin totals of the second. The iterative proportional fitting procedure now determines the set of data cells c'_{ij} that approximate both marginal totals. This is accomplished by executing two steps several times:

1. The c_{ij} are adjusted to equal the marginal column totals by first dividing each cell by the actual sum of the column of cells, and then by multiplication by the marginal column total:

$$c^1_{ij} = c_{ij} \frac{t_{\cdot j}}{c_{\cdot j}} \tag{4.2}$$

2. Each row of now column-adjusted cells are then adjusted to equal the marginal row totals.

$$c_{ij}^{2} = c_{ij}^{1} \frac{t_{i\cdot}}{c_{i\cdot}^{1}} \tag{4.3}$$

The two steps are repeated either until the selected level of convergence is reached or if no more changes take place. See figure 4.1 for an example by HUNSINGER. The figure shows a given matrix of seed cells and marginal totals in 4.1(a). In 4.1(b), the data cells that are obtained after IPF are shown. The data cells now more closely correspond to the marginal totals while still reflecting the seed distribution. For example, the first row's second cell remains the highest of that row, so does the third row's first cell.

The procedure can be analogically extended to work with more than two variables. Note that the sum of the marginal row totals and the sum of the marginal column totals must be equal and in order to avoid a division by zero, all of the marginal cell values must not be zero. The obtained data cells will in most cases never exactly sum to all of the predetermined marginal totals. If seed cells are set to zero, no adjustment will be made to those. A workaround is to choose a comparatively small value to those (e. g. 0.001 if all of the other marginal cells are whole numbers) [46].

1	2	1	5
3	5	5	15
6	2	2	8
11	9	8	

(a) before IPF

$\xrightarrow{\text{IPF}}$

1.51	2.31	1.18	5
4.20	5.35	5.45	15
5.28	1.34	1.37	8
11	9	8	

(b) after IPF

Figure 4.1: Iterative proportional fitting (IPF) demonstration example [46]. The matrix in (a) shows the seed cells representing a given distribution that does not correspond to the marginal totals aside. (b) then shows the changes made to the table after several iterations.

4.3 Sources for spatial input data

It is certainly not much of a surprise that apart from transport-specific empirical data, transport models also incorporate spatial data. In recent years, both data quantity and quality have increased significantly, also due to the development of modern methodologies in form of airborne and satellite sensors that provide large sets of information on urban areas [10]. Those and accompanying innovations in data acquisition, data preparation and data processing open up new ways for the integration of spatial data into transport models.

Nagel and Marchal outline the development process of microscopic simulations and identify it as a straight forward process. At first, a simulation substrate must be created and populated with agents, which are then supplied with rules defining their behavior [66]. In the most simple version, the substrate represents the physical space as a projection on a two-dimensional plane. Alternatively, a designated model can be used as substrate, allowing a more complex description of the simulation area. That in combination with a terrain or building model allows to resolve the substrate in 3D. In the following, spatial data sources are introduced, which, once taken together, represent a detailed simulation substrate.

4.3.1 Land-use data

Of the vast potpourri of private and governmental providers for spatial data, the most important one for Germany is perhaps the Federal Agency for Cartography and Geodesy (*Bundesamt für Kartographie und Geodäsie*, BKG), a national organization whose purpose is the provision of reference and spatial data for administrative and scientific needs [21]. Besides other data products, the agency maintains a digital landscape model (DLM) and a digital terrain model (DGM).

The first is a two-dimensional digital description of geo-spatial objects for entire Germany at an aspired resolution of ± 3 meters. It corresponds roughly to a 1:25 000 scale topographic map [5] and is part of ATKIS, the official German topographical cartographic information system (*Amtliches Topographisch-Kartographisches Informationssystem*) [104]. An example showing an extract of the data is given in figure 4.2. The latest version of the DLM is accessible on the Web at www.geodatenzentrum.de through a java applet as well as machine readable XML web services. Alternatively, a local copy can be obtained in form of ESRI shape files. Typical structures that are represented are roads, railway tracks, rivers and others. In total, the catalog of objects

distinguishes between 170 separate object types, divided into the areas 'settlement', 'transport', 'vegetation', 'water bodies', 'relief' and 'territories'. For most objects, their position and contour is on record, whereas some objects are stored in a simplified form as a single point or line. The latter is often used to represent objects smaller than 3 meters in diameter, although the minimum object size varies depending on the object type [5]. Additionally, DLM-objects are described by attributes that further specify the object. For instance, attributes are used to specify a building's function, i. e. whether it is a church, a school, a community center, city hall, a train station etc. Attributes can provide very deep knowledge on the object that they describe, reaching as far as, for example, the number of floors that a building has. The revision cycle for ATKIS data is one year for the road network and five years for all other object types. Thus in practice, obtaining updated versions regularly seems appropriate [98, 5]. A precise definition for each object type can be found in [105]. Besides data acquisition and distribution, the function of the BKG is also to monitor the quality of the data. A recent examination is presented in [5]. A drawback is that at this time entire Germany is not covered at the same level of detail, as the data collection process is not yet completed. Some regions are more affected by this than others. Especially highly specialized attributes, like for instance a power plant's main source of energy (nuclear, coal, gas, solar, ...) are likely candidates.

Of special interest to the field of transportation is the description of land-use areas by the DLM. Water bodies or woodland, for example, should clearly show low to zero road-traffic congestion as opposed to urban areas. The significance of land-use is that it can be considered for the disaggregation of population data (see chapter 6.2).

The digital landscape model contains vector data. As this is quite often constructed from remotely sensed data, it should be subject to a rigorous accuracy assessment [86, 5]. With respect to the BKG's aim to assess and improve the quality of the data, the data is assumed to be the most reliable data source for Germany. Equivalent foreign landscape models are also accessible for most western countries through national databases, and the work on a standardized European landscape model is in process.

4.3.2 Buildings and postal addresses

Another dataset that is maintained by the Federal Agency for Cartography and Geodesy (BKG) is a listing of address coordinates for entire Germany. For every postal address, a coordinate is given that describes the position in space that the ad-

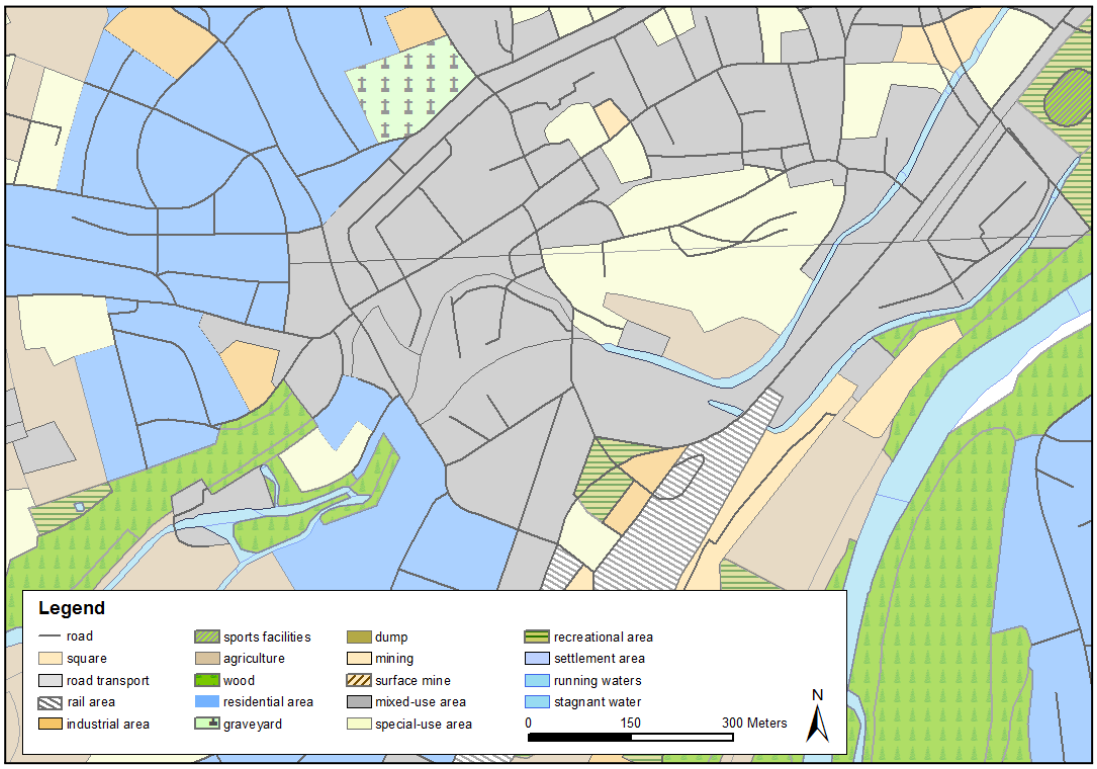


Figure 4.2: Example extract of the digital landscape model (DLM). Source: *Digitales Landschaftsmodel (DLM)* of *Bundesamt für Kartographie und Geodäsie (BKG)*.

dress refers to. Usually, some solid, built structure can be found there, which has lead to the synonym 'house address'. An extract of said record set is displayed in figure 4.3, with emphasize on its high spatial resolution. In the figure, each point represents a coordinate, with the four land-use patterns 'industrial', 'residential', 'mixed-use' and 'special-use' shown underneath.

The house coordinate record set is maintained as a separate, stand-alone dataset. Its origin are the digital representations of buildings that are collected by each federal state. The BKG as a national authority then obtains this data. Despite the high precision, the data set is at this point not yet spatially inclusive and comprehensive. It is therefore supplemented with address records provided by the Infas Institute for Applied Social Sciences (www.infas.de). As of June 2009, a total of 1 450 387 address records are added. The supplementation is carried out by BKG, with the resulting set of address records being then made available as a separate dataset called *Georeferenzierte Adressdaten – Bund* (GAB). The supplementation process is laid out in [20].

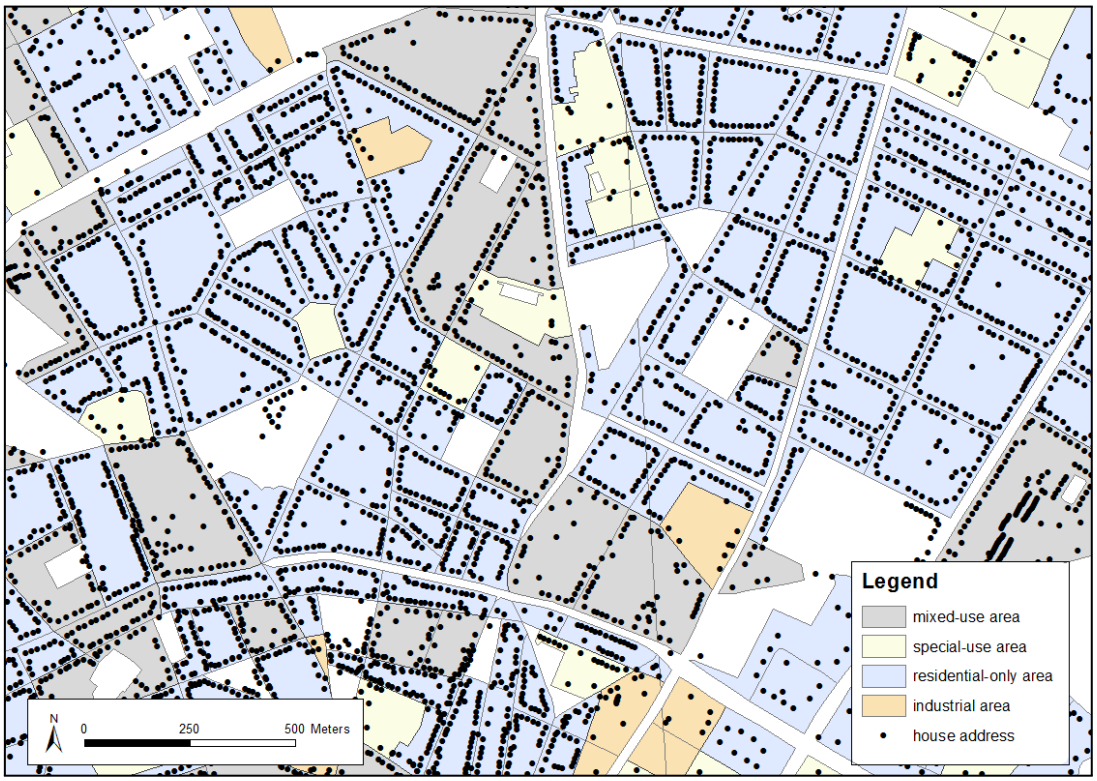


Figure 4.3: Example extract of the house address data set showing a Berlin suburb. Each spot marks a building that is linked to a designated address record with street, house number and postal code. The colored shapes underneath represent land-use areas that are taken from the land-use model DLM. Source: *Georeferenzierte Adressdaten – Bund (GAB)* and *Digitales Landschaftsmodell (DLM)* of *Bundesamt für Kartographie und Geodäsie (BKG)*.

4.3.3 Spatial zoning systems

A third data type that is important for the creation of simulation substrates are zoning systems. Zoning systems break space down into various geographical zones and serve two purposes. One is to connect statistical data to space. Statistical data is often cut into geographical zones, each zone being represented by a unique key. The corresponding zoning system then defines polygons that represent the geographical area behind each key. Another is to aggregate the sets of individual trips into zone-based origin-destination (OD) matrices. Several zoning plans that are common for Germany and one European counterpart are exemplified in the following:

- **Official Municipality Keys** (KGS) (for the German *Kreisgemeindeschlüssel*). The Federal Republic of Germany is divided into 12 385 official municipalities with

different authority. Each of them is identified by an eight-digit number, which is generated by the concatenation of predefined codes for state, government district, city district, and municipality. Some municipalities are unincorporated areas, which are, though not politically independent, nevertheless part of the municipality partitioning system [96]. Because of its length of eight digits, the partitioning system is sometimes referred to as KGS 8.

The system's hierarchical structure allows zones with different levels of detail. For example, when only the first two digits are incorporated, no more than the sixteen states are distinguished. A level of detail that is often found in national statistics and empirical studies features the first five digits (KGS 5). In this case, the level of detail reaches as far as urban areas (in a district-free city) or districts (in a city with districts).

It is important to note that the partitioning system changes when new districts are introduced, or current ones are split or combined through land reforms. One major change was caused by the reunification of East and West Germany in 1989. However, the partitioning system has since been subject to change many times. For example, in 2002 a total of 13 222 areas were defined, while by 2008 the number had decreased to 12 300. Therefore, when connecting two separate data sets, their reference date must be the same. The current KGS zones as well as the ones of recent years can be obtained as ESRI Shape files from the Federal Agency for Cartography and Geodesy (*Bundesamt für Kartographie und Geodäsie*, BKG).

- **Statistical areas** have been defined by the Federal Statistical Office (*Statistisches Bundesamt*). This partitioning system extends the one previously listed by adding 12 257 more districts. These districts are defined in urban areas where a more detailed zoning is pursued. The city of Berlin, for example, is defined by one single KGS 8 key, and is then further divided into 287 statistical areas.

- **Nomenclature of Territorial Units for Statistics** (NUTS – for the French *nomenclature des unités territoriales statistiques*). This nomenclature is similar to the KGS system, but covers all member states of the European Union. It is organized in several levels: NUTS-0, NUTS-1, NUTS-2, NUTS-3 [32]. The first level differentiates the actual EU member country, and for each, a hierarchy of three more levels is established. Note that the subdivisions in some levels do not always correspond to the administrative divisions within the country.

- **Postal codes**. The system of postal codes in Germany consists of approximately eight thousand cells, each one of them identified by a unique five-digit code (which can feature a leading zero). The city of Berlin, for example, is split into 190 postal code areas.

- **Traffic cells**. Certain zoning systems split regions into zones that are optimized for transport-specific considerations. For example, PTV AG, a corporation that develops and dispatches traffic simulation software, maintains a zoning system for this purpose. About 7 000 cells are defined, out of which around 250 represent European countries, whereas all others represent Germany [76]. Several other zoning systems dedicated to transportation exist, as for example the regional zoning system that is maintained by the Statistical Institute of Berlin-Brandenburg (*Amt für Statistik Berlin-Brandenburg*) for the area of Berlin, Germany. The latter will be referred to in the demonstration example in chapter 6.

The need for interpreting several zoning systems comes with the integration of a variety of different input data sources. So does the empirical survey *Kraftfahrzeugverkehr in Deutschland* (KiD) from section 4.2.1 specify waypoints by geographic coordinates and postal code, whereas the population data from section 4.2.3 refers to municipalities, statistical areas and traffic cells.

4.4 Conclusion

Starting from the theory of activity-based personal transport models and by deciding against modeling the human decision process due to the problems that are associated with it, the conceptual basis for the present approach is established. The fact that trip chain data is in fact available and that this data contains attributes that also appear in standardized spatial data suggests to bring both together. The trend towards microscopic simulation systems for personal transport demand generation and the increasing availability of precise spatial data suggests to develop a microscopic simulation system. In this context, standardized cartographic landscape models, which are either available or under construction for most western countries, provide a qualified substrate for simulation and, with the help of standardized spatial zoning systems, empirical data can be linked to it.

5 The proposed methodology

The feasibility analysis in chapter 4 leads to the proposed methodology which will be presented in this chapter. It will summarize the procedure in section 5.1 and then provide a detailed description subsequently in the ongoing sections: section 5.2 will introduce the components that represent the model system and section 5.3 will provide formulas which specify the total traffic volume that the model system generates. Section 5.4 then presents the methodology's principal approach followed by analyses towards a computerized implementation in section 5.5 and 5.6.

5.1 Outline of the methodology

The separate model steps are drafted in consecutive order in figure 5.1 and can be summarized as follows:

1. To begin with, a virtual world is created first, which involves sets of synthetic firms and private households. Firms are described by a set of attributes and supplied with a geographical allocation in space, which is done at the level of individual buildings. The locations of buildings are coded by their postal address consisting of street, house number and postal code. Households are in the same way as firms located by street, house number and postal code.

2. Next, the traffic volume that is generated by a firm is determined based on the firm's attribution. For every attribute combination that occurs in the synthetic world, a mobility ratio is derived from empirical data in order to reflect the number of agents that operate on behalf of firms of that type.

3. Knowing the number of agents that perform trips, the next task is to define where they go. This is accomplished by the emulation of typical behavior patterns in the form of logbook trajectories. Just as in the previous step, logbook trajectories that fit to a firm's attribution are used. For implementing the approach into software, the concept of template logbooks will be introduced.

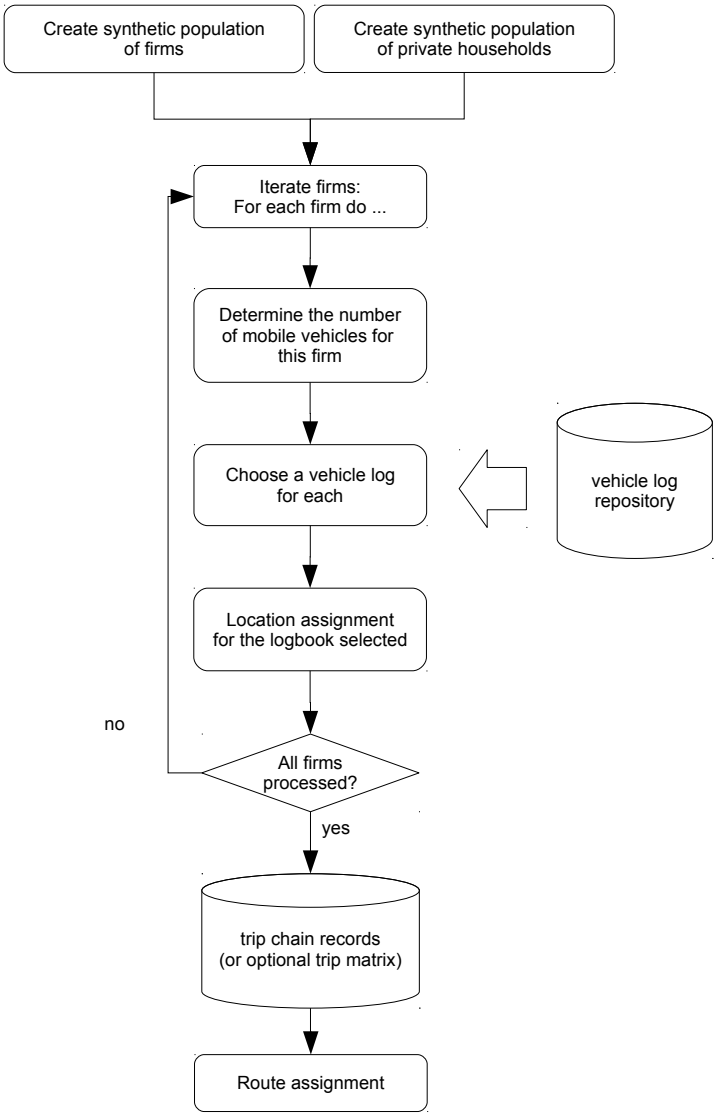


Figure 5.1: Outline of the proposed simulation procedure.

4. By processing every business in the virtual world, all trips on behalf of these firms are generated resulting in the region's total transport demand of those firms.

The methodology produces detailed output statistics at both the aggregate and the disaggregate levels. Most importantly, it produces trip chains as space and time trajectories for each vehicle. The trip chains include the location of activities, the purpose of each trip, the vehicle type that was used, and others. The approach also produces aggregated statistics over all vehicles such as the total number of trips, the total distance traveled, the average number of stops per trip chain, and so on. Statistics can be obtained for any subset of vehicles classified on any of the characteristics that the template logbooks provide. And due to the coding of waypoints as spatial coordinates, statistics can be spatially refined, for example obtaining the transport demand between two regions.

5.2 Model components

The model M can be defined as

$$M = (C, B, H, R, I) \tag{5.1}$$

where C is a set of locations in the form of geo coordinates, B is a set of businesses, H is a set of private households, R represents a repository of template logbooks and I is a set of instructions for location assignment. C, B, and H together form the virtual world, whose components are presented hereafter in section 5.2.1. The repository R is discussed subsequently in section 5.2.2. I represents a system of parameters and restrictions for location assignment that are discussed in detail in section 5.4 to 5.6.

5.2.1 Virtual world

Locations are represented through spatial coordinates. Origin and destination of trips will later refer to elements of C:

$$C = \{c_1 \dots c_n\} \tag{5.2}$$

where n is the number of locations.

The set B of businesses represents all businesses that exist in the virtual world:

$$B = \{b_1 \dots b_n\} \tag{5.3}$$

where n is the number of businesses.

Accordingly, H is the set of private households

$$H = \{h_1 \dots h_n\} \tag{5.4}$$

where n is the number of households.

Further, a business $b_i \in B$ is characterized by a number of attributes:

$$b_i \to (a_1 \dots a_m)_i \tag{5.5}$$

where $a_1 \dots a_m$ are attribute values for the attributes in A_B, the set of the attributes that describe firms.

For the demonstration example in chapter 6, the attribution will be chosen to consist of the economic sector a_{sector} that best reflects the firm's activities and the firm's size a_{size}:

$$a_{\text{sector}} \colon B \to \mathcal{E} \tag{5.6}$$

and

$$a_{\text{size}} \colon B \to \mathcal{S} \tag{5.7}$$

where $\mathcal{E}$ is the set of economic sectors (e. g. manufacturing, construction, education) and $\mathcal{S}$ a set of classes representing a business's size measured in number of employees (e. g. 1-9, 10-49, 50-99, and so on).

For economic activities, several standardized classification schemes exist. The most prominent one for Germany is WZ 2003 by 'Bundesamt für Statistik', the federal statistical office of Germany. NACE (for the French *nomenclature statistique des activités économiques dans la Communauté européenne* [31]) is a classification scheme on European level, to which WZ 2003 is compatible. Another classification scheme

on international level is ISIC (International Standard Industrial Classification) by the United Nations [95]. A choice between these can be made with regard to the input data sources. For the empirical survey KiD (see section 4.2.1) as well as the population data (see section 4.2.3), this is WZ 2003.

Furthermore, private households and businesses are positioned in space through assignment to locations. Therefore, a mapping is created between the set of locations C and private households and businesses respectively:

$$f_b \colon B \to C \tag{5.8}$$

$$f_h \colon H \to C \tag{5.9}$$

Note that both mappings are not meant to be injective, because several private households and businesses are able to share the same location. Nor are they surjective, because the intention is not to connect every location to at least one business or private household. What methodology should be chosen for connecting businesses and private households to locations depends on the existence and availability of suitable data sources for the study area. In chapter 6.2 on page 74, a procedure is shown which generates mapping tables for f_b and f_h on the basis of land-use information.

The methodology for location assignment which will be presented in chapter 5.4 very much relies on the shaping of the synthetic world. In order for the proposed location assignment algorithm to work properly, the synthetic world must be defined in adequate size. The meaning of "adequate" can be best described by distinguishing between two separate areas:

1. Firstly, the area for which traffic is to be simulated. It will be referred to as the study area. If, for example, a model is to be designed that simulates the traffic of a particular city, this city's border line would set the limit for the study area.

2. Secondly, the area for which synthetic world data exists, that is the virtual world.

The two areas must be at least identical in size, so that the first does not exceed the second. For location choice not to be biased by areas that are not defined, the study area should be a smaller extract from what is covered by the virtual world.

5.2.2 Logbook repository

The movement of vehicles is simulated by emulating behavior patterns that are defined in template logbooks. All template logbooks l_i together form the logbook repository R:

$$R = \{l_1 \dots l_n\} \tag{5.10}$$

where n is the number of template logbooks.

A template logbook $l_i \in R$ contains trips for one vehicle for a given period of time and is further specified through attributes. Rather than keeping logbooks for every possible combination pair that theoretically exist based on the attribution, template logbooks must be kept only if businesses of that type exist in the synthetic world.

Universally expressed, each template logbook is characterized by a tuple of attributes:

$$l_i \rightarrow (a_1 \dots a_m)_i \tag{5.11}$$

where $l_i \in R$ is a template logbook and $a_1 \dots a_m$ are the attribute values for the attributes in A_L that describe the template logbook.

For the particular implementation that is demonstrated in chapter 6, the two attributes economic sector a_{sector} and firm size a_{size} are chosen to specify the business type to which the logbook is linked:

$$a_{\text{sector}} \colon R \rightarrow \mathcal{E} \tag{5.12}$$

and

$$a_{\text{size}} \colon R \rightarrow \mathcal{S} \tag{5.13}$$

where $\mathcal{E}$ is the set of economic sectors (manufacturing, construction, education) and $\mathcal{S}$ a set of classes representing a firm's size.

Besides information on the firm to which a logbook is connected, the attribution can further describe the vehicle. For example, the vehicle type a_{vtype} may be defined as:

$$a_{\text{vtype}} \colon R \to \mathcal{V} \tag{5.14}$$

where $\mathcal{V}$ is the set of vehicle types that are differentiated (e. g. gasoline and diesel).

Further, logbooks contain trips. Thus, a logbook l_i must consist of one or more trips t_{ij}:

$$l_i \to \{t_{i1} \dots t_{in}\} \tag{5.15}$$

where n is the number of trips of l_i. Typical examples of various template logbooks are given in figure 5.2.

Trips are characterized by their distance covered $d(t_{ij})$ as well as their origin and destination. The trip's distance is stated as a metric distance function (direct airline distance, routed distance or travel time). Choosing the airline distance simplifies the mathematical calculations during location assignment, since standard geometric operations then apply: all destinations in reach lie on a circle with the radius being the airline distance from the trip's origin. Other distance functions would not lead to circles, but to isochrones, that is non-regular shapes which are more computationally intensive to process. For regions in which transport infrastructure and natural barriers such as rivers and mountains cause strong effects on the distance/time ratio,

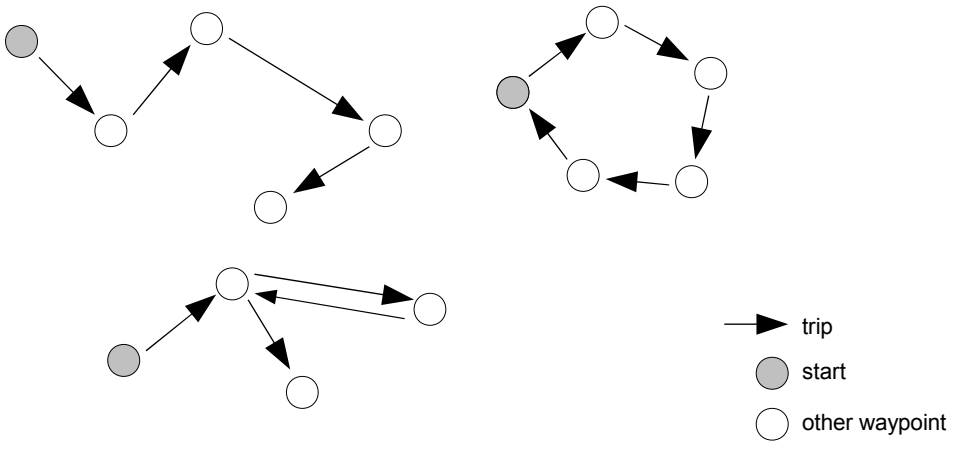


Figure 5.2: Template logbooks examples.

implementing travel time or routed distance should be considered. One example is the Switzerland area, for which AXHAUSEN AND HURNI carried out an in-depth analysis and produced graphical representations in the form of time maps that emphasize the difference between spatial distance and travel time [3]. For the demonstration example in chapter 6, the trip distances are expressed by the air-line distance, assuming that Berlin has few major barriers to motorized traffic. Expressing the distance function as the actual travel time allows to account for congestion effects by other vehicles in the road network. By executing several simulation runs in which the travel time obtained in the previous run is passed into the current run, the effects can be passed on to location assignment.

In order to specify the type of a trip's origin and destination, a set of waypoint types $\mathcal{W}$ is introduced:

$$\mathcal{W} = \{w_1 \dots w_n\} \tag{5.16}$$

where n is the number of waypoint types that are distinguished. Waypoint types w_i provide information about the function of a waypoint, for example whether the destination is an external company, a private household, or perhaps a terminal facility. As a requirement, one and only one element of $\mathcal{W}$ must represent a vehicle's home site, that is the business site to which a vehicle is associated:

$$\exists_1 w \in \mathcal{W}, \text{ with } w \text{ is 'home site'} \tag{5.17}$$

A trip's origin and destination can now be specified by having them point to one element of $\mathcal{W}$:

$$w_o \colon T \to \mathcal{W} \tag{5.18}$$

and

$$w_d \colon T \to \mathcal{W} \tag{5.19}$$

where T is the set of all trips as the sum of all logbooks:

$$T = \bigcup_{l \in R} T_l \tag{5.20}$$

and T_l the set of trips of logbook l.

Further attributes can be supplied if provided by the empirical basis. The empirical survey *Kraftfahrzeugverkehr in Deutschland* (KiD), for example, specifies among others the trip purpose, the number of passengers and the type of load (if goods are carried with). In order to incorporate trip purposes (e. g. going to work, shopping, recreation), a set $\mathcal{P}$ of purpose types is introduced:

$$\mathcal{P} = \{p_1 \dots p_n\} \tag{5.21}$$

with

$$p\colon T \to \mathcal{P} \tag{5.22}$$

There exist logbooks that contain trips with different purposes. An example is that of a service technician visiting a number of customers before noon, thus generating for the moment commercial passenger transport trips. If the technician then heads home for lunch, a trip with a private purpose is generated. The technician then might continue their customer visits after lunch causing again the generation of commercial passenger transport trips. Information on a trip's purpose can be used to filter trips accordingly after simulation, e. g. if only trips of specific type should be passed on to route assignment. The latter can be used to prevent conflicts that arise when the outputs of several autonomous demand models are merged, e. g. if a separate private transport demand model is used to generate private trips.

By summing up the distances of the trips t_{ij}, the total distance $d(l_i)$ of a logbook l_i can be expressed:

$$d(l_i) = \sum_{j=1}^{k} d(t_{ij}) \tag{5.23}$$

where k is the number of trips of l_i.

Accordingly, a logbook's remaining distance $d^r(l_i, x)$ of a logbook l_i after a given trip number x is:

$$d^r(l_i, x) = \sum_{j=x+1}^{k} d(t_{ij}) \tag{5.24}$$

where l_i is a template logbook, x is a given trip's number in chronological order, k is the total number of trips of logbook l_i, and $d(t_{ij})$ is the distance of trip j of logbook l_i.

Finally, information on a logbook's shape is retained. The idea behind is demonstrated in figure 5.3, in which two logbooks that consist of four trips each are displayed. Although the two logbooks share equivalent trip distances, they are different in shape. In other words, the tour pattern determines a vehicle's action radius. Therefore, a variable z_i is introduced that computes the distances between a logbook's starting point and each waypoint, in the following referred to as a waypoint's distance-to-startsite. Figure 5.4 displays a logbook record and the corresponding values z_i. Let x_0 be the starting location of a logbook and x_i all the waypoints in chronological order. Then z_i can be defined as

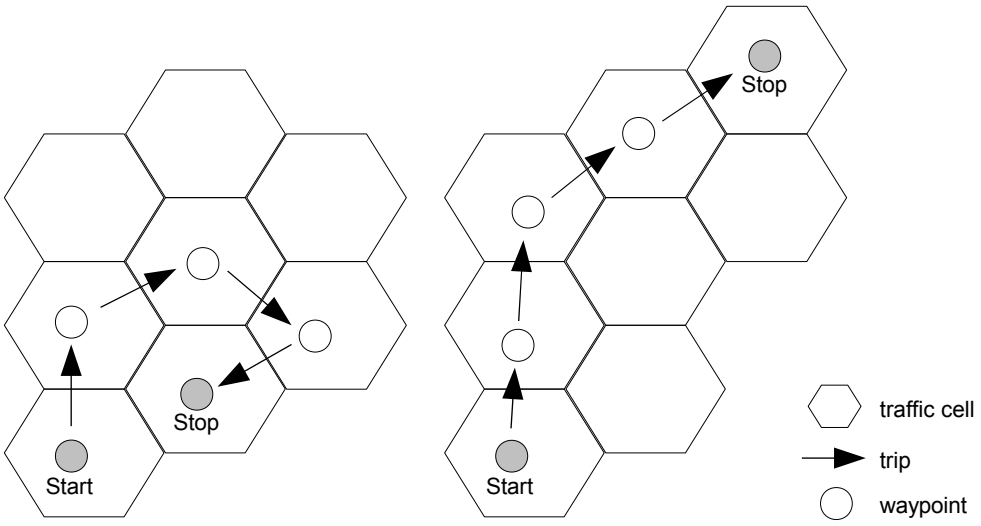


Figure 5.3: Two logbooks of different shape but equivalent trip distances. The figure shows two logbook records. Although trip distances and therefore total distance are equal in both graphs, the actual pattern is different, causing one logbook to cross distant traffic zones while the other remains near its starting point.

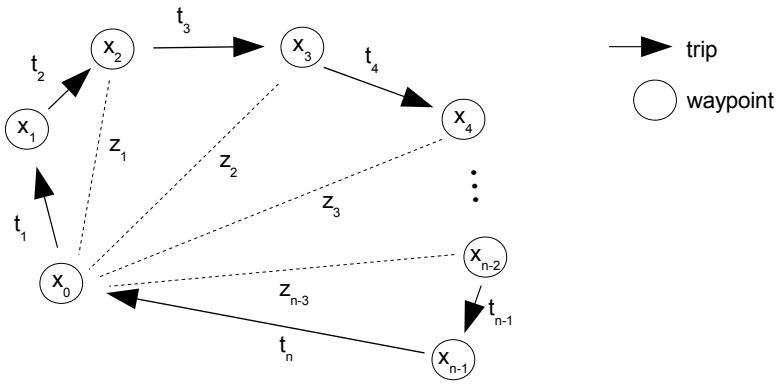


Figure 5.4: A waypoint's distance-to-home (z_i). The figure shows a logbook record with corresponding distances z_i, n = total number of trips.

$$z_i = d(x_{i+1}, x_0) \qquad \text{for } i = 1 \ldots n-2 \tag{5.25}$$

where n is the number of trips and $d(a, b)$ the euclidean distance between a and b:

$$d(a, b) = \sqrt{(a_x - b_x)^2 + (a_y - b_y)^2} \tag{5.26}$$

Summarizing, a template logbook l_i is described through the tupel

$$l_i = (a_1 \ldots a_n, t_{i1} \ldots t_{im})_i \tag{5.27}$$

where $a_1 \ldots a_n$ are the attributes that characterize the logbook (e. g. firm size, economic sector, vehicle type) and $t_{i1} \ldots t_{im}$ are the trips of l_i.

Each trip is described by the tuple

$$t_{ij} = (a_1 \ldots a_n, w_o, w_d, d(t_{ij}), z(t_{ij}))_{ij} \tag{5.28}$$

where $a_1 \ldots a_n$ are the attributes that characterize the trip (e. g. the trip's purpose $p(t_{ij})$), w_o and w_d are the types of origin and destination respectively, $d(t_{ij})$ is the distance, and $z(t_{ij})$ is the distance between the trip's destination and the origin of the logbook's first trip (distance-to-startsite).

5.3 Total traffic volume and logbook to business assignment

The hypothetical – as will become clear later – total number of vehicle kilometers v that are generated during simulation is the sum of the vehicle kilometers produced by all businesses:

$$v = \sum_{b \in B} v(b) \tag{5.29}$$

where $v(b)$ is the traffic volume produced by business b.

In the simplest version, $v(b)$ is determined by two components. The first is a mobility ratio $m(b)$ that determines the number of vehicles operating on behalf of business b. The second one is the distance covered by each vehicle. For each vehicle, the latter is determined through a logbook selected from the relation R_b which links firms to template logbooks:

$$R_b \colon B \to R \tag{5.30}$$

The relation $R_b \subset R$ states the subset of logbooks from which logbooks are taken for a given business b. One way to formulate R_b is to have the firm's attribution meet the one of the template logbook:

$$R_b = \{l \in R \mid a_1(l) = a_1(b) \wedge \ldots \wedge a_i(l) = a_i(b)\} \text{ , for } \forall a_i \in A_B \cap A_L \tag{5.31}$$

where A_B is the set of attributes that describe firms and A_L is the set of attributes that describe template logbooks.

For the current implementation that is described in chapter 6, R_b contains all logbooks that share the economic sector and the firm size of b:

$$R_b = \{l \in R \mid a_{\text{sector}}(l) = a_{\text{sector}}(b) \wedge a_{\text{size}}(l) = a_{\text{size}}(b)\} \tag{5.32}$$

An alternative is to let R_b express a mapping backed by other empirical findings that tell what kind of travel behavior is typical for firms according to their characteristics.

Having R_b contain the set of logbooks from which can be chosen for a given business b and having further the mobility ratio $m(b)$ specify the number of vehicles that operate

on behalf of b, next, a logbook is chosen for each vehicle from R_b either by random or, where applicable, by using a weight function. Accordingly, the traffic volume $v(b)$ produced by business b is

$$v(b) = \sum_{l \in R'_b} d(l) \tag{5.33}$$

where R'_b is the set of logbooks that were chosen for business b, with $|R'_b| = m(b)$.

The mobility ratio $m(b)$ must be either derived from empirical data or be determined through estimation models. Section 6.4 will exemplify how ratios were derived from the empirical survey on commercial passenger transport that was anticipated prior in chapter 4.2.2. Depending on the empirical data at hand, the single mobility ratio $m(b)$ can be replaced by a set of ratios $m_i(b)$, for instance to distinguish between several vehicle types (number of trucks, number of cars, ...) as visualized in figure 5.5.

So far, the formulas feature raw distances as defined in the template logbooks. Later in chapter 6.3, logbooks are derived from the survey *Kraftfahrzeugverkehr in Deutschland* (KiD) and thus consist of trips that actually took place in a specific regional context. In order to be able to apply a logbook to other regional settings, a strict interpretation of the distances is not wanted. In consequence, the distance is allowed to vary at a given rate. The outcome is a new total distance $\tilde{v}$:

$$\tilde{v} = \sum_{b \in B} \tilde{v}(b) = \sum_{b \in B} \sum_{l \in R'_b} \tilde{d}(l) \tag{5.34}$$

Tolerance is realized by allowing a generated trip's distance to differ from the template's distance. The degree to which they can differ is limited by a parameter ϵ that specifies how far the trip distance $\tilde{d}(l)$ in the emulation can differ from the template logbook's trip distance $d(l)$. In order to prevent unfavorable effects caused by the adjustment, $\tilde{v} = v$ is aimed at by choosing rather small values for ϵ. The influence of the parameter ϵ will be described in full detail in the following section 5.4, which introduces a set of constraints for location choice.

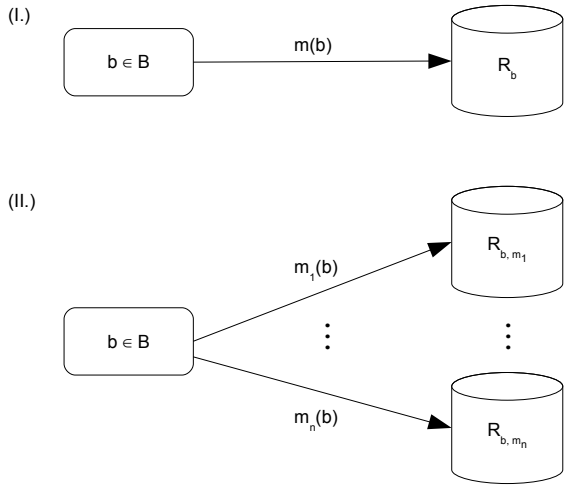


Figure 5.5: One or multiple mobility ratios per firm.

5.4 Constraint definition

For location assignment, vehicles are moved within the synthetic world as indicated by template logbooks that follow the definition given in section 5.2.2. As presented ahead in section 5.3, for each business the number of mobile vehicles that operate on behalf it must be determined – either by estimation models or from empirical data. Each vehicle is then modeled by selecting from the logbook repository a template logbook that corresponds to the business' attribution. Generated trips should closely reflect the given template logbook, although it most certainly originated from an area other than that of the business to which it is applied.

This section explains the proposed location assignment procedure. At first, two constraints are introduced that together define potential locations for a trip's destination based on waypoint type in section 5.4.1 and on trip distance in section 5.4.2. Then in section 5.4.3, a sorting criterion is introduced that ranks the resulting locations according to their similarity to the original template logbook's shape.

5.4.1 The waypoint type constraint c_1

The first constraint c_1 requires the types of waypoints to be equivalent to the ones of the template logbook in successive order in order to maintain the activity sequence that is defined by it. For example, if a template logbook states that the first trip points to a private household, then a trip to a private household should be generated. If the template logbook states next that the second trip points to a business site, a business should be chosen as destination for trip number two, and so on. While the constraint in fact may be modeled in many ways (e. g. keeping the first and the last trip, but allowing to modify the order of the intermediate trips), it is defined here to "clone" the template logbook to generate a close copy of the original. The main argument for high similarity between template and generated logbooks versus allowing for modification of the template logbook's characteristics is the few scientific evidence on the trip planning process. The constraint can be adapted to account for the new knowledge in future.

At first, a search pattern s_w is defined for each waypoint type $w \in \mathcal{W}$ that describes the typical characteristics of that type:

$$s_w \rightarrow \text{restriction } 1 \wedge \text{restriction } 2 \wedge \cdots \wedge \text{restriction } n, \text{ for all } w_i \in \mathcal{W} \tag{5.35}$$

where $\mathcal{W}$ is the set of waypoint types and restriction 1 to n can represent any condition that can be tested on the available data. With the filter, for every waypoint type, out of all locations that are part of the synthetic world, only those are interpreted as potential trip destinations that meet the search pattern of the waypoint type. In consequence, all others are not considered for further processing.

Search patterns can contain several components in order to capitalize multiple information sets that are provided in the synthetic world. For example, a location's attribution in combination with land-use data of where it is located can be taken into account. The following search pattern defines locations of type 'shipping agency' to be located at any location where a business of economic sector 'Transport, storage and communication' is situated and which is further located in an industrial area according to local land-use data:

$$\begin{aligned} C_{\text{shipping agency}} = \{c \in C \mid c \text{ “is business”} \\ \wedge\, a_{\text{sector}}(c) = \text{“transport, storage and communication”} \\ \wedge\, a_{\text{land-use}}(c) = \text{“industrial”}\} \end{aligned} \tag{5.36}$$

In consequence, locations that do not meet these criteria do not come into question. Further constraint definitions can follow any detail level available in the used data. Employees of communication service providers, for example, could be directed with a higher chance to households with high speed internet plans. Likewise, salesmen could be directed more frequently to areas of high purchasing power or transportation hubs might be defined to be in the near vicinity of highway ramps.

Universally expressed, for a given trip t_{ij} the constraint c_1 can be stated as

$$c_1 \colon C_{t_{ij}}^{c_1} = \{c \in C_w \mid w = w_d(t_{ij})\}\text{, for } t_{ij} \in T \text{ and } w \in \mathcal{W} \tag{5.37}$$

where $\mathcal{W}$ the set of waypoint types, T the set of trips as the sum of all logbooks, t_{ij} the j'th trip of the i'th logbook and C_w the set of locations that correspond to the waypoint type definition for type w.

5.4.2 The trip length constraint c_2

The second constraint c_2 requires a trip's length to meet the distance that is specified in the template logbook. If, for instance, a trip must cover 10 kilometers according to

a template logbook, then only objects at that distance should be considered. Hence, the constraint can be formally expressed as

$$C_{t_{ij}}^{c_2} = \{c \in C \mid d(s,c) = d(t_{ij})\} \tag{5.38}$$

where C is the set of locations, $d(t_{ij})$ the distance of the i'th trip of the template logbook l_j, and $s \in C$ the trip's starting location.

However, this constraint needs to be weakened, since some tolerance must be allowed for. One reason is that computational precision requires this measure, because otherwise distance would be interpreted in a scale of centimeters or millimeters. Such precision is not requested, since the constraint's precision should not outperform the precision of the input data. Another reason is that a far too strict interpretation of the distance attribute will lead to large numbers of template logbooks not being reproducible to the synthetic world. Thus, reasonable tolerance is needed to avoid the set C^{c_2} to be empty in too many cases. For the two reasons, the constraint is expanded by a parameter ϵ that defines to what extend the length of a trip may be altered during simulation:

$$c_2 \colon C_{t_{ij},\epsilon}^{c_2} = \{c \in C \mid d(s,c) \geq (d(t_{ij}) - \epsilon) \wedge d(s,c) \leq (d(t_{ij}) + \epsilon)\} \tag{5.39}$$

High values will make possible solutions more likely and smaller values will lead to a higher degree of similarity between generated trips and their template logbook. Epsilon can be defined based on the actual trip distance, such as to allow the distance in the simulation to vary by ± 10 per cent of the original trip distance. For the demonstration example that is later presented in chapter 6, a fixed value of $\epsilon = 50$ meters was chosen to test the approach with a small value for all trips. The ideal value for epsilon is defined by the optimum between the two factors (1) replication precision and (2) replication success rate. The first demands that generated trips be close to the original templates and the second requires large numbers of logbooks be successfully applied to the synthetic world, e. g. to have less than x logbooks for which the assignment failed. In consequence, epsilon can be determined by iterative simulation runs until the optimal value is found.

5.4.3 The sorting criterion using z

The intersection of both sets, being the set of objects containing only objects that adhere to both constraints, summarizes the requirements from section 5.4.1 and section 5.4.2:

$$C^{c_1,c_2} = C^{c_1} \cap C^{c_2} \tag{5.40}$$

In the likely case of this subset containing more than one object, an order is created between them by emphasizing the template logbook's spatial shape based on the value z, which was introduced in chapter 5.2.2. Recall that z reflects the distance from a waypoint to the starting location of a logbook's first trip. By following the steps of the algorithm which is presented in section 5.5, all objects show the same distance to this location if a logbook's first trip is to be generated, because z and the trip's distance represent the same line segment in this particular case. Here, random order is used. For all other trips (from trip number two upwards), the elements of C^{c_1,c_2} are sorted by how well their z values correspond to the template logbook:

$$\min \left| z(t_{ij}) - z^{'}(t_{ij}) \right| \tag{5.41}$$

where $z(t_{ij})$ is the distance-to-start for trip t_{ij} of a template logbook and $z^{'}(t_{ij})$ is the distance-to-start of a location in C^{c_1,c_2}.

Where appropriate, the degree at which a waypoint's distance-to-start can vary from that of the original template logbook can be limited:

$$C^{c_z}_{t_{ij},\gamma} = \left\{ c \in C \mid \left| z(t_{ij}) - z^{'}(t_{ij}) \right| \leq \gamma \right\} \tag{5.42}$$

where C is the set of locations, t_{ij} is the i'th trip of the template logbook l_j, and γ is the maximum acceptable difference.

5.5 Template logbook preparation

Now that the constraints for the location assignment procedure are defined, next, the template logbooks are adapted for the upcoming steps. First, formal requirements that template logbooks must comply to in order to be used for the approach are presented in section 5.5.1. In section 5.5.2, conversion rules are then presented which transform logbooks into an alternative representational form. This form allows logbooks to be processed by the two algorithms for trip chain generation that are presented subsequently in section 5.6.

5.5.1 Formal requirements

Special attention is given to trips that either originate from or point to a vehicle's home site. This is because, unlike other waypoint types, the business's location is unambiguously defined in the synthetic world and as such remains fixed during location assignment. Thus this location shall serve as a pivot point around which all trips are then placed, as illustrated in figure 5.6.

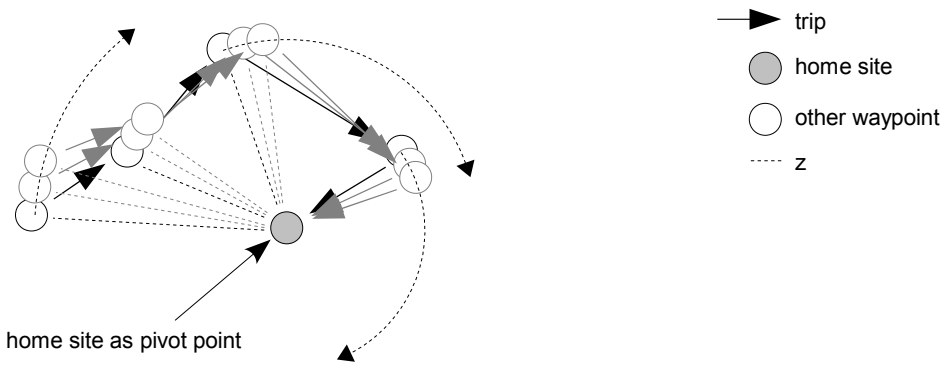


Figure 5.6: The concept for location assignment. The figure shows the application of a template logbook to the virtual world, here depicted by rotating a logbook around the home site until a suitable position is found. That is, the locations must correspond to those of the template logbook by their type, the distances must be equivalent and the tour's geometrical pattern must approximate the template's shape. The home site is taken as pivot point because its location is unambiguously defined in the synthetic world.

Hence, logbooks are classified by how many waypoints of type 'home site' they contain:

1. Logbooks with no stop at home site.
2. Logbooks with one stop at home site.
3. Logbooks with two or more stops at home site.

Figure 5.7 shows an example for each group. Logbooks of type two and three can be used for simulation, while logbooks of type one (no stop at home site) are problematic, since no pivot point is defined for such.

In case of a large number of logbooks being of type one, several possibilities exist:

1. First, with the homesite not appearing in a template logbook but a firm's position being the only fix point defined in the synthetic world, no hint is available that tells where a logbook should be placed. Thus, a "straight forward" solution is to disperse them randomly in space.
2. Another solution is to extend the virtual world in a way so as to include other items that do occur in the template logbook. For example, logbooks that lack waypoints of type 'home site' probably have trips that return home, that is they point to the private household where the driver lives. If relationships between

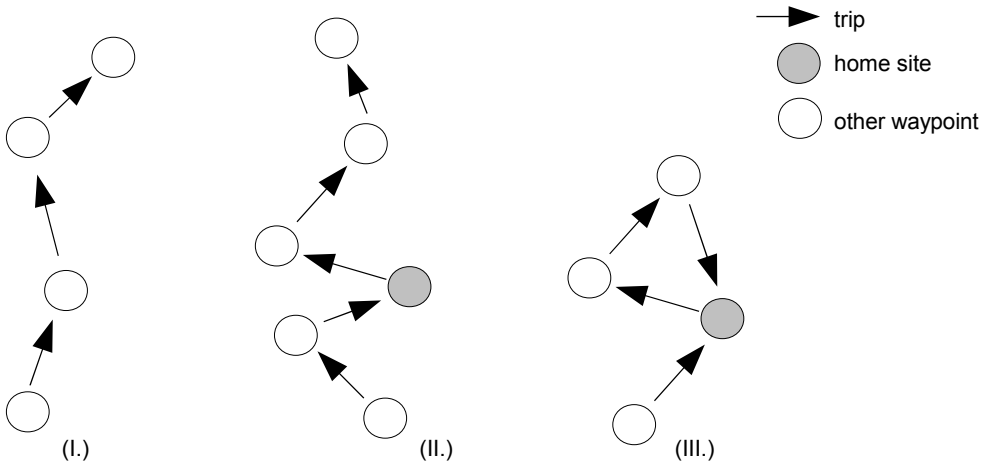


Figure 5.7: Logbook categorization for location assignment. The figure shows three logbook records. Logbook (1) does not stop at home site at all. Logbook (2) stops once, while logbook (3) stops twice.

citizens and their work places were part of the synthetic world, the driver's private household could be used as pivot point instead.

3. A third alternative is to account for information on how far from the home site a vehicle was operating, if supplied by the empirical basis. The empirical survey *Kraftfahrzeugverkehr in Deutschland* (KiD), for example, allows to determine the distances from a vehicle's home site to every waypoint, and in consequence defining the radius around which the vehicle was operating.

During the demonstration example that is presented in chapter 6, affected logbooks were selected in 16 % of all cases. Since many of these were vehicles operating outside of the study area as their distance-to-homesite exceeds the study area's actual size, none of the solutions from above is applied and logbooks of type one (no stop at home site) are not further processed.

5.5.2 Transformation procedure

All remaining logbooks are converted into a simplified form in order to prepare them for the upcoming steps. To begin with, logbooks with two or more waypoints of type 'home site' (group three) are split into a number of smaller ones. A cut is done each time the home site is reached, as displayed in figure 5.8. Logbooks with precisely one stop at home site (group two) are split up as well, unless the stop at the home site is either located at the source or the final destination.

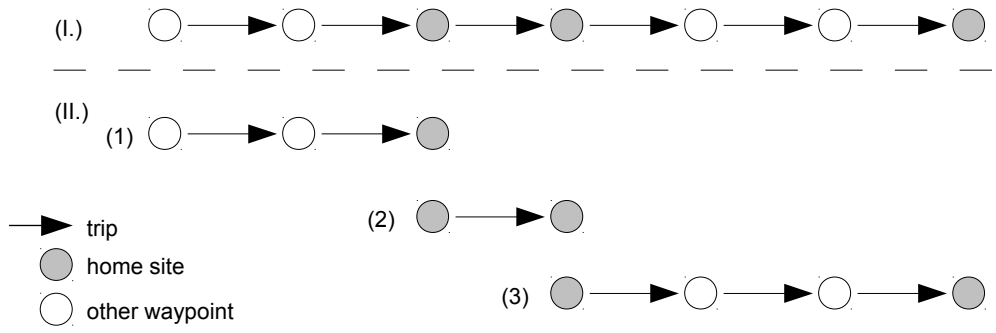


Figure 5.8: Splitting logbooks. A cut is done each time the vehicle stops at its home site. In the example, splitting the logbook from the top (I.) returns the three sub-logbooks shown beneath (II.).

From now on, all logbooks contain no more than two stops at 'home site', and if a waypoint is of type 'home site', it must be at either a logbook's first or last position. Thus, logbooks can now be broken down into the following grouping:

1. Logbooks that start (but do not stop) at a waypoint of type 'home site'.
2. Logbooks that stop (but do not start) at a waypoint of type 'home site'.
3. Logbooks that start and stop at waypoints of type 'home site'. Since precisely one home site is defined for every business in the synthetic world, logbooks of this kind start and stop at the same point. Therefore such logbooks form a closed trip chain in which, after processing every trip, a vehicle finds itself back in the same position from which it started.

By reversing the order of waypoints – meaning that the first trip will become the last, the last trip will become the first, and so on – , group two can be eliminated since reversing will place all logbooks of group two into group one. The reversal is demonstrated in figure 5.9.

The principal tenor is to only alter a logbook's formal representation without changing much of the actual information that it contains, which is why the original order of waypoints and the trip directions need to be preserved. Therefore, when splitting a template logbook, the resulting sub-logbooks are grouped together by organizing them into a sequence at which they occurred prior the conversion process as displayed in figure 5.10. I. e. a single logbook is replaced by a collection of sub-logbooks with a known sequence and in addition, if a trip's direction is changed, the direction prior change is noted. That is, the set of logbooks contained in a logbook collection resemble the original logbook of that collection.

At this stage, the preparation of logbooks is completed. Figure 5.11 on page 60 summarizes the steps. All logbooks, or all sub-logbooks for those that were split, now either

1. do not form a closed trip chain (only source is of type 'home site', all other waypoints are of other types) or
2. form a closed trip chain (source and final destination are of type 'home site', all other waypoints are of other types).

This further implies that at this point the source of any logbook (or sub-logbook) is of type 'home site'.

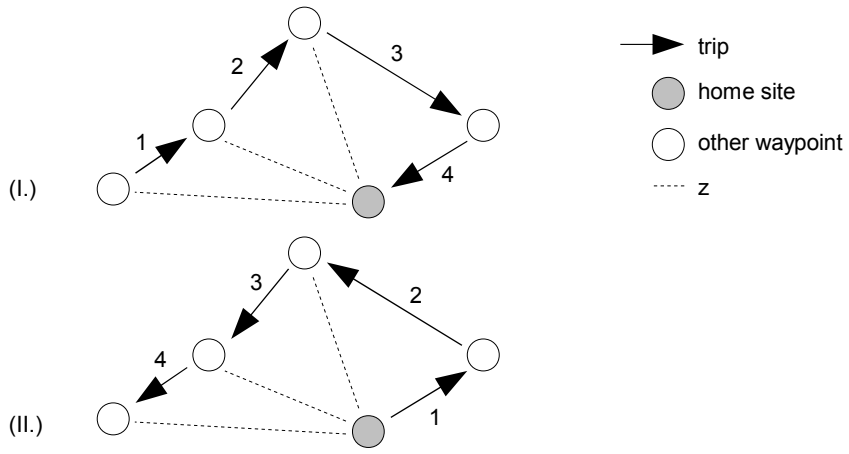


Figure 5.9: Reversing a logbook's trip order during location assignment. The figure shows the original logbook (1) at the top, with its final destination of type 'home site'. Logbook (2) shows a logbook which resembles the one above in reversed trip order.

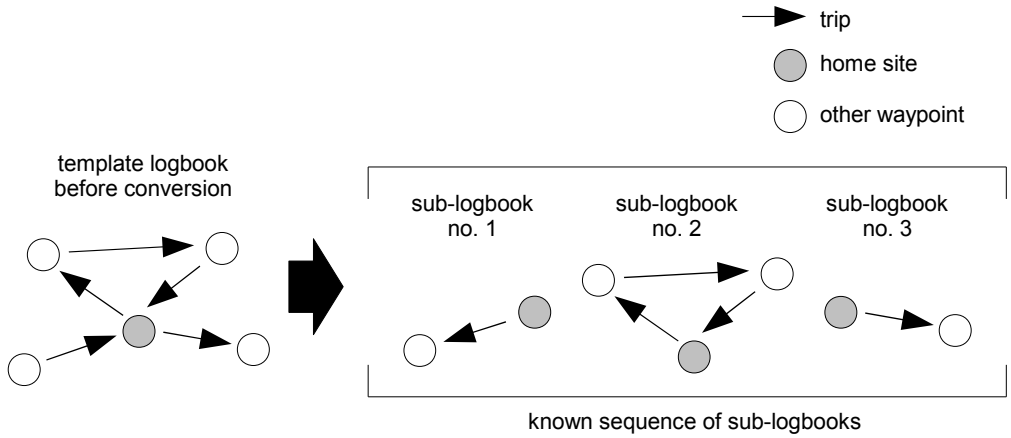


Figure 5.10: Sequencing sub-logbooks into logbook collections in order to preserve the logbook's original structure prior conversion. The three logbooks in sequence resemble the original template logbook. A trip's direction, if changed, is also kept on record, here the case for the first trip.

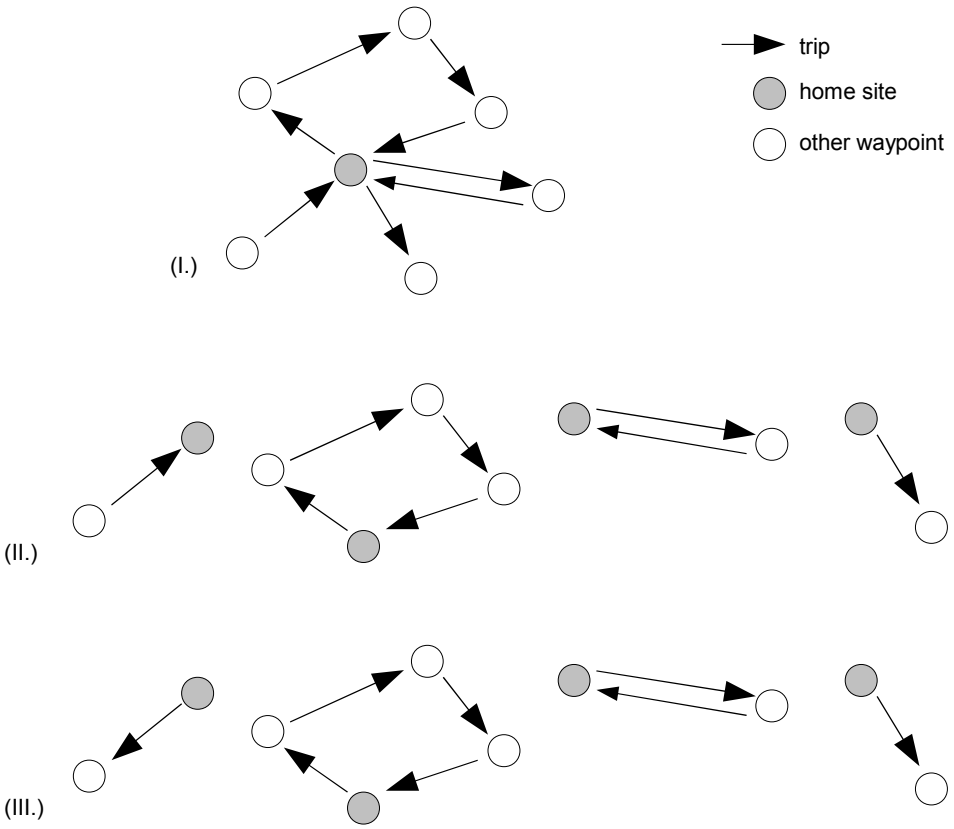


Figure 5.11: Logbook transformation for location assignment. The figure shows the initial logbook (1) at the top, which is then split into four sub logbooks (2). The first of these is then put into reversed order (3), with the effect that all logbooks now start at the home site.

5.6 Algorithmic solution

Now that the logbooks adhere to the above classification, two separate algorithms are introduced – namely one for each group. The algorithms ultimately generate synthetical trip chains based on the preprocessed template logbooks from section 5.5 whilst meeting the set of constraints from section 5.4. The two algorithms have in common that they start at a firm's home site and then move on trip by trip. The first is presented in section 5.6.1 and works with logbooks of group one. The second is an extended version of the first for logbooks of group two (section 5.6.2). For the

exceptional case that trips go beyond the designated area that is to be covered by the model, a substitute procedure is presented in section 5.6.3.

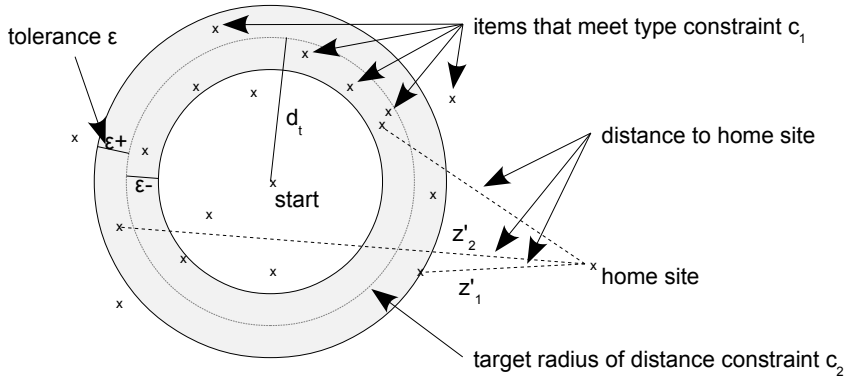
5.6.1 Algorithm for logbooks of group one (open trip chains)

An algorithm for logbooks of group one (only source is of type 'home site', all other waypoints are of other types) will be shown in the following. As already anticipated prior in section 5.5.2, special attention is given to trips that either originate from or point to a vehicle's home site, since, unlike other waypoint types, the business's location is unambiguously defined in the synthetic world. Since the business that the algorithm is going to process is known and since all businesses are positioned in space, a precise coordinate can be retrieved at which the home site of the business is located.

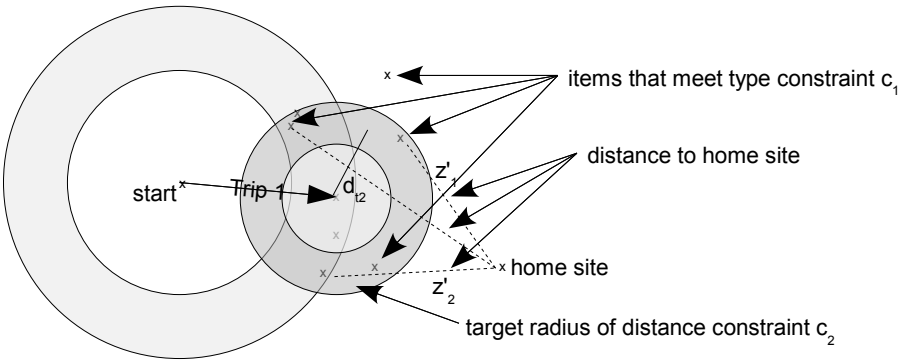
The algorithm starts by selecting potential destinations for the logbook's first trip. Only objects from the model world are considered which (1) correspond to the trip's destination type and (2) are at the given distance. The procedure is illustrated in figure 5.12(a), where the distance is interpreted as the direct airline distance between the two waypoints. In consequence, the target area of the distance constraint c_2 is a circle.

All objects that meet the two criteria are ordered according to the sorting criterion z. Whether it is the hierarchy's first object that is chosen or the hierarchy's second object and so on depends on whether any subsequent trips exist that must be generated. If so, the following trip is processed in the same way as the first one, with the hierarchy's first choice as starting location. Figure 5.12(b) demonstrates the procedure for trip number two. The procedure is repeated until all other remaining trips are processed. While the methodology is the same for every trip, location type and distance stem from the template logbook and can therefore vary from trip to trip. The recursion terminates once all trips of the template logbook have been generated. If at some point a trip cannot be assigned because no object exists that meets restrictions c_1 and c_2, a backtracking procedure is applied, that is the previous waypoint is discarded and the next object in line is chosen instead. Starting from this point, the procedure is then continued until a valid solution is found. If it turns out that another trip cannot be assigned, the algorithm once again steps back and selects the hierarchy's next object, and so on.

As a matter of fact, the algorithm terminates also if all possibilities failed and thus no possible solution exists that allows the emulation of the template logbook inside



(a) The location assignment procedure. The figure demonstrates how a trip's destination is chosen among the objects of the synthetic world. All marks represent objects that adhere to the first constraint c_1. With the trip's source location x_{start} as center, a circle is then drawn with the trip's distance d as radius, so as to limit the set of objects to those that lay on it (constraint c_2). Note that the circle's thickness is expanded by the parameter ϵ, so as to allow for tolerance. All remaining objects are then sorted according to how well their distance to home site reflects the template logbook. The list's first object, i.e. the object that reflects the template logbook best, is selected. If forthcoming trips (as displayed in (b)) cannot be assigned successfully, the next object in line is chosen and so on.



(b) Location assignment for the generation for trip number two. Based on the destination that was selected for the first trip, a circle is drawn with the second trip's distance as radius. The marks display all objects that meet type constraint c_1 of trip two.

Figure 5.12: The location assignment procedure.

the synthetic world. For such cases, an alternative procedure must be defined. Here, three possibilities exist:

1. First, a different logbook than the one that failed can be selected. If it turns out that the logbook which was chosen at first is not applicable to the region in which the business of the current simulation is located, then a different logbook might be applicable instead. Considering that the logbook which failed was chosen at random from the logbook repository, selecting another logbook with the same attribution can be an alternative.

2. The previous solution has the drawback that it affects the drawing which otherwise happens at random. Therefore, an alternative is to ease the restrictions c_1 and c_2, making the assignment process more flexible and thus a valid solution more likely. For example, a greater tolerance value ϵ increases the likelihood of successfully applying the same logbook (demonstrated in figure 5.13 on the following page). This approach has the advantage that the changes apply to all logbooks, causing every logbook to be processed with the same rules.

3. A third alternative is to change the distances of the trips in a way that the altered logbook becomes applicable to the current region. That should preferably be done by shifting distance from one trip to another, so as not to change the logbook's total distance. A disadvantage is that in this case, the template logbook is merely altered "to make it fit" in a pragmatic way rather than being informed by knowledge.

However, during the demonstration example further on in chapter 6, the proposed location assignment algorithm required such interventions only in 2.3 % of all cases. See chapter 6.5 for a detailed analysis.

If trip distances are not expressed by the airline distance, but by the travel time or the routed distance, the target area of the distance constraint c_2 would then be an isochrone. As such, distance functions other than airline distance can be considered at the cost of greater run time. The increase in runtime is due to two reasons. One is that processing isochrones, as they are non-regular shapes, is a more complex operation and hence takes more processing power than circle-based operations. However, experience indicates that with modern Geographic Information Systems (GIS), if supplied with sufficient hardware, the increase can be kept reasonable low. Another reason is that the actual isochrones may not be known and hence must be generated, leading to

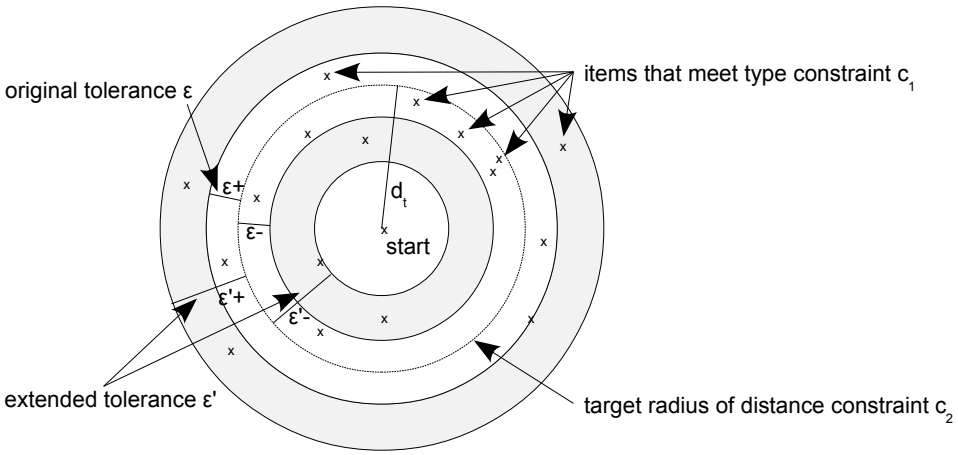


Figure 5.13: Allowing for spatial tolerance in location assignment. The two stripes in dark gray mark where potential waypoints can additionally occur if the tolerance value ϵ is increased.

higher complexity. Opposed to airline distance, travel time is not constant and varies by factors such as the time of the day, road congestion and others. A compromise is to generate static isochrones beforehand or to draw from empirical data.

5.6.2 Extended algorithm for logbooks of group two (closed trip chains)

The assignment of logbooks that form a closed trip chain, that is both source and final destination are of type 'home site', utilizes a third constraint c_3 which restricts the set of objects that come into consideration further. The use of the constraint c_3 does not affect the algorithm's output, but significantly enhances the algorithm's runtime, because many locations that do not lead to a final solution are excluded at an early stage. In consequence, the set of items that are passed on to the next recursion step is minimized. The constraint considers the fact that, in case of closed trip chains, the vehicle must come back to its initial starting location. Thus, how far a waypoint can be from a logbook's initial starting point is limited by how much distance is left by all remaining trips.

The constraint c_3 can be formally written as

$$c_3 : C_{l,i}^{c_3} = \{c \in C \mid d(l_S, c) \leq d^r(l, i) + \epsilon(k - i)\} \tag{5.43}$$

with l being the template logbook of which locations are being assigned, l_S the log-

book's starting location, $d^r(l,i)$ the remaining distance of logbook l after trip i, and k the total number of trips of l.

The intersection of all three sets must be built accordingly:

$$C^{c_1,c_2,c_3} = C^{c_1} \cap C^{c_2} \cap C^{c_3} \tag{5.44}$$

Figure 5.14 demonstrates the usage of the additional constraint. For simplicity reasons, the parameter ϵ is left out. Note that when implementing the constraints into software, the processing speed can be optimized further by executing the constraints in appropriate order.

If a template logbook was split into a number of sub-logbooks beforehand, each is processed separately. For sub-logbooks of group one (open trip chains), the algorithm for logbooks of that group is chosen and vice versa, for sub-logbooks of group two, the algorithm for closed trip chains is applied. By processing all sub-logbooks, a trip chain is obtained that resembles the characteristics of the original template logbook, causing the generated synthetic trips to equal the original template logbook as a whole. In particular, the number of waypoints, the order of waypoints and the waypoint types correspond to the original logbook and in addition, the trip distances between the waypoints and hence also the total distance correspond to the template logbook (with the exception of the tolerance that is allowed for by the parameter epsilon). And furthermore, trips that were reversed are set back to their original form, ensuring that the final output is not altered by the preparation steps from section 5.5.

5.6.3 Trips transcending the synthetic world

The two algorithms presented so far also terminate if a trip's distance exceeds the synthetic world. That is the case if the synthetic world's definition is too narrow for the trip to be placed into it. In test runs for the Berlin area, this case occured quite frequently when the model world was only defined within the city limits. Therefore, this section is dedicated to trips that exceed the synthetic world's definition range.

At first glance, vehicles may not be tracked any further once they exit the study area. This is because, by leaving the study area behind, their trips are no longer focus of the simulation. Nonetheless, computing time is spent to continue the simulation of such vehicles in order to monitor whether they reenter. Because if they do so, their succeeding trips should be considered. However by matter of fact, the procedure presented so far in the sections 5.4 to 5.6 cannot be applied any longer once the

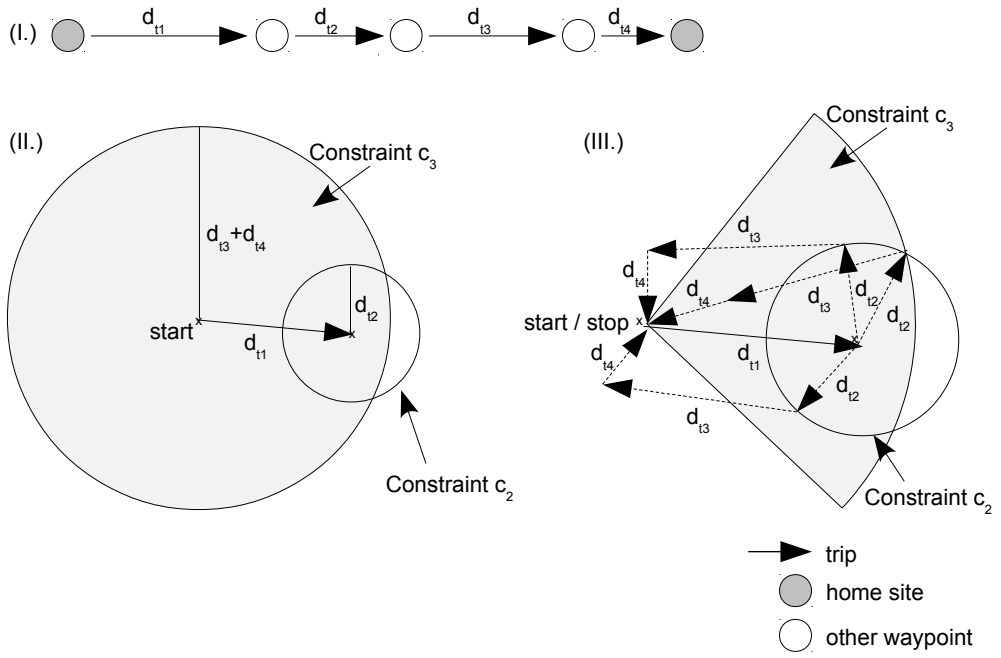


Figure 5.14: Extended location assignment for closed trip chains (logbooks whose start and final stop is of type 'home site'). The template logbook that is used for demonstration is displayed in the top-left corner (I.). Figure (II.) shows the effect of the additional constraint c_3 which is represented by the large circle. In regular location assignment, i.e. location assignment for open trip chains, all locations lying on the small circle, whose radius is the trip's distance, come into consideration. In the case of closed trip chains, the additional constraint restricts the locations to all those that are inside the large circle. Its radius is the distance that remains by the following trips. Taken together, all locations that lie on the small circle and find themselves inside the large circle are passed on to the next recursion step. Figure (III.) finally gives examples of possible tour patterns.

synthetic world's border is crossed, since the necessary data is not defined beyond it. Instead, an adapted version must be developed based on data that is available, specifically (1) a template logbook's trip distance and (2) a waypoint's distance-to-homesite (the distance between the waypoint and the business on whose behalf a vehicle operates). In place of the constraint definitions c_1 and c_2, the set of waypoints is now defined through the constraint

$$c_o = \{x \in S \mid d(s,x) = d(t_{ij}) \wedge d(h,x) = z(t_{ij})\} \tag{5.45}$$

The constraint limits the number of suitable locations from space S to all those that show the same trip distance and the same distance-to-homesite. The latter is measured from a business' homesite h, while the trip's distance $d(s, x)$ is measured starting from the previous waypoint s. In a graphical illustration as displayed in figure 5.15, this is equivalent to spanning two circles, their radiuses being the trip's distance and the distance-to-homesite. The two circles intersect in either one, two, or zero points, thus causing the set c_o to contain that many elements. In the case that both circles do not intersect, only the trip's distance is considered by selecting a finite number of points lying on the distance circle to be passed on to the next recursion step. In consequence, c_o contains as many elements as the number of selected points. The procedure is applied for all remaining waypoints either until all waypoints are processed or the vehicle reenters the study area, from which point on regular location assignment is again applied. As a consequence of the adapted location choice algorithm, trips beyond the virtual world no longer originate or point to defined locations such as businesses or private households. Since locations do not exist outside the virtual world, origin and destination are directly coded in geographical coordinates of space S.

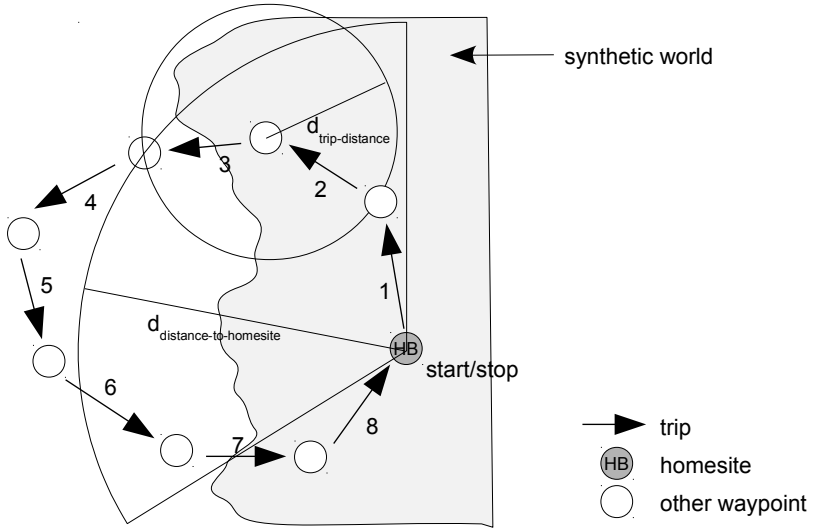


Figure 5.15: Trip assignment in areas where no virtual world data is defined. A template logbook whose third trip leaves the virtual world is shown. The destination of this trip is then generated by calculating the intersection of two circles. One circle has the trip's distance as radius, the radius of the other circle reflects the trip's distance-to-homesite. The process is continued until the vehicle reenters.

5.7 Summary of the proposed methodology

The presented methodology generates synthetic trip chains by mapping empirical trip chain data to a regional context. The methodology satisfies most of the requirements of section 4.1, with the exception that modeling the decision process is abandoned. For demonstration and evaluation purposes, the model system has been implemented into software, which will be the focus of chapter 6.

6 Example demonstration and evaluation

This chapter tests the methodology on a real-world example by simulating one average work day for the city of Berlin, Germany. The chapter presents implementation aspects and assesses the simulation results. The empirical implementation has three purposes: The first is to demonstrate the applicability with respect to input data requirements. Second, the demonstration is to show that the simulation will terminate in time, even for study areas of considerable size. And the third is to test whether the template logbooks that are extracted from the empirical survey *Kraftfahrzeugverkehr in Deutschland* (KiD) can be effectively mapped to the study area.

The city of Berlin is chosen as study area because of several reasons. Firstly with respect to its size, as Berlin is a large urban area. The *Amt für Statistik Berlin-Brandenburg*, the official statistical office for the region, reports the city to cover 892 square kilometers and being home to 3 444 000 inhabitants as of may 31, 2010 [1]. The area's size is a challenge. So will the house data set by the Federal Agency for Cartography and Geodesy (BKG) supply roughly 1.1 million houses to the simulation. Successfully applying the algorithms to this area will show that, in terms of processing speed, they can also be applied to other smaller areas. Second, Berlin has been the subject of research in previous projects (examples are [6, 11, 12]), allowing to re-use some of the data that was used for these projects. All data sources that are used for the simulation are summarized in table 6.1. Furthermore, the German Aerospace Center (DLR), where this thesis is written, is linked to local authorities with its research site in Berlin-Adlershof, making the simulation of this region a direct interest. And finally, the demand for commercial transport for the Berlin region has been assessed in previous works, allowing a comparison of the simulation results with this data.

The chapter is organized into a number of sections. As a prerequisite, the software that runs the simulation is developed in section 6.1. Next, the relevant input data is collected and fed to the system. At first, the virtual world is created in section 6.2 by supplying private households and firms to the simulation. Thereafter, the template logbook repository is filled with logbooks from the empirical survey *Kraftfahrzeugverkehr in Deutschland* (KiD) in section 6.3. The mobility ratios that specify the number of

Table 6.1: The input data for the demonstration example.

Input data	year
Private household population [68]	2005*
Firm population [60]	2005*
Land-use data [104, 105, 5]	2007
House coordinates [20]	2008
Survey *Kraftfahrzeugverkehr in Deutschland* [100]	2001/2002
Survey *Dienstleistungsverkehr in industriellen Wertschöpfungsketten* [43]	2007

* generated with data of different years, the primary one as of 2005.

vehicles that operate for each firm is then determined in section 6.4. And section 6.5 finally presents the actual simulation run. The simulation parameters are set and the results are presented, including a test against third-party aggregates.

6.1 Software implementation

6.1.1 Spatial database systems

With regard to the computational implementation, the use of a geographic information system (GIS) is an obvious choice for processing spatial data. Yet, there is also data to be processed without spatial context. Thus the perfect solution to embed the model into is a system capable of handling both data with and without spatial context. One solution for this purpose are relational database systems, since they have become the standard tool for storing and processing data. A drawback is that the speed of such systems can be insufficient for the specific needs of large-scale micro simulations [66]. In consequence, individual data structures are developed for specific model designs that provide faster data access. Some efforts even go further and investigate into the use of dedicated hardware, see for example Strippgen and Nagel [92] for a system that makes use of a computer's graphical processing unit (GPU). Yet for the present implementation, the use of a relational database system is favored because of several reasons. First, their ease of use allows to easily test different algorithms with a minimum of programming effort. They further support the rapid extraction of useful statistics, allowing one to analyze the processed data at any stage of the computation process. Furthermore, modern database systems are capable of processing large quantities of spatial data, allowing to concentrate on the programming logic rather than on implementation issues without major performance loss. Fourth, predefined interfaces for in- and export allow to quickly add new data to a simulation. And most importantly, the method's constraints that were defined in chapter 5 can be directly translated into the Structured Query Language (SQL), the language for the management of data in relational databases.

Among the most popular spatial database systems, with no claim to be complete, are Oracle Spatial [71], PostgreSQL with PostGIS extension [78], Microsoft SQL Server [63], and MySQL Spatial Extension [65], out of which PostgreSQL is chosen. PostGIS is an additional module that extends the standard PostgreSQL database by adding the ability to efficiently process indexes on spatial data types. In addition, predefined stored procedures that come with PostGIS provide many standard geo calculations such as distance, area, or spatial intersection. Spatial data can be imported and exported as ESRI Shape files, a common proprietary geospatial vector data format.

The main argument for choosing PostGIS is that it follows the OpenGIS Simple Features Specification for SQL, a standard SQL schema supporting storage, retrieval,

query and update of geospatial data [70]. Being developed by the Open Geospatial Consortium (OGC), an international industry consortium consisting of companies, government agencies and universities, this data model can be migrated to other database systems [69]. Further, PostGIS offers the benefits of free software. And finally, PostGIS is supported by a large user community that produced numerous postings on the web, a criterion that distinguished it clearly from MySQL when the work on this thesis started.

6.1.2 Reference coordinate systems

Despite of the actual database system that is used, spatial data typically relates to a reference coordinate system. The 'World Geodetic System 1984' (WGS 84) and the 'Gauss-Krüger coordinate system' (GK 3) for the Germany, Austria, and Finland area are among the most common ones. For the city of Berlin, the 'Soldner coordinate system' is also used. At the time of writing, the *Bundesamt für Kartographie und Geodäsie* (BKG), a federal authority for geographic data in Germany, used GK 3 as their main reference system. Following their choice, the present project is set up in GK 3 as well and if geometries happened to be in a different reference system, they were transformed into GK 3. This is because erroneous or misleading results can occur when geometries are kept in different reference systems. Note that the BKG is currently in the process of establishing the triple A data model, an improved version of the existing one that requires to move to the 'European Terrestrial Reference System 1989' (ETRS 89) [13]. Hence, ETRS 89 may be chosen for future implementations.

For identifying a geometry's reference system, PostGIS uses 'Spatial Reference System Identifiers' (SRIDs), a system for distinguishing between the different reference systems by unique numbers. At the time of writing, 3 162 predefined reference systems were distinguished by default. Those that were used during data collection and preparation for the present demonstration example are displayed in table 6.2.

6.1.3 Data model and system design

With the key aspect of this thesis being the conceptional methodology, only a short summary of the actual implementation is given here. The ease of use of relational database systems, their open connectivity, the ability to rapidly develop and test code and the ability to express this methodology's constraints in SQL make the use of a relational database system the preferred choice for implementation. A relational data

Table 6.2: Spatial Reference System Identifiers (SRID)

SRID	Description
31467	Gauß-Krüger-Abbildung, 3. Meridianstreifen (Mittelmeridian 9°) [GK 3]
4326	World Geodetic System 1984 [WGS 84]
3068	Deutsches Hauptdreiecksnetz (DHDN) / Soldner Berlin
4258	European Terrestrial Reference System 1989 [ETRS 89]

Source: www.spatialreference.org [22]

model is designed that integrates the digital landscape model (DLM) from section 4.3.1 into the database system and that further contains tables for the method's objects such as firms, households, and locations. A Java client then connects to the database. Its main purpose is the execution of the simulation loop structure that iterates over the simulation entities. In principle, all code may be implemented inside the database management system in the form of stored procedures using PL/SQL, making any additional code outside the database dispensable, but Java was chosen for ease of use. In order to benefit from the database's capabilities of efficiently processing spatial queries, all spatial calculations are processed by the database system.

6.2 Population data synthesis

The synthetic urban space is supplied with private households and firms based on two separate synthetic population files. At this point, the population files contain individual firms and individual private households at the level of geographical zones. According to the definition in section 5.2.1, households and firms need to be mapped to locations for the present approach. Therefore, a methodology is presented in this section for the disaggregation of entities that are coded at the level of zones into the level of separate buildings. The process is informed by consulting the cartographic data on land-use and on building locations that was previously introduced in section 4.3. The section is split into three parts. To begin with, four land-use types are discussed in section 6.2.1. The actual disaggregation algorithm is then presented for private households in section 6.2.2 and for firms thereafter in section 6.2.3.

6.2.1 Using land-use data for disaggregation

Land-use can be seen as an important aspect for many physical processes taking place on the earth surface [5]. Using it for transport modeling is advantageous with respect to one of the basic aspects of modeling. Generally, a model must simplify the complicated real world to a more simple scale and in this sense elide certain features of reality. Yet there are things which we tend to know more about than others, with land-use belonging to the category that we know very well. Land-use can be analyzed easily, since it is not hidden and can be directly observed from the real world without any elaborate methodologies necessary to "unveil" how things really are. This especially accounts for designated, man-made land-use districts such as enterprise zones and residential areas. Since we create them ourselves, no complex data acquisition and analysis must be carried out as is needed for land-use patterns that are the result of natural processes. Another important argument is that land-use is generally not restricted by data privacy requirements and thus available in high precision for most areas, since it does not relate to individuals. For the above reasons, land-use is considered to be a helpful supplement for the disaggregation of population data.

Land-use data can be obtained from the digital landscape model (DLM) that was introduced earlier in section 4.3.1. While the landscape model contains a large number of designated land-use types, specifically four are of interest for the disaggregation process, as firms and households are typically located within these types and by definition are rarely located in other areas [105]. The four areas are:

1. **Residential area**: An area characterized by buildings that serve housing and living purposes in first place. Places for shopping, culture, religion, handcraft and health services can be found as well.

2. **Industrial area**: An area that is mainly used for industrial and commercial purposes. This can include shopping malls, storage depots, large-scale commercial operations, and trade fair areas.

3. **Mixed use area**: An area characterized by buildings not dominated by any specific use. These are mostly found in rural areas and inner city districts that combine retail, housing, and public services.

4. **Area for special use**: An area characterized by buildings that serve a particular distinct function. Examples are, among others, public administration, health and social services, education, science, culture, recreation and the military.

Figure 6.1 shows how the four land-use types distribute for Berlin and the surrounding area. Note that by considering the four land-use types described above, the remaining land-use types that typically do not contain firms nor households (waterbody, farmland, nature reserve, and so on) are actually taken into account through exclusion.

6.2.2 Private household population synthesis

The synthetic population file was designed at the Institute of Transport Research (VF) at the German Aerospace Center (DLR) as part of a research project that was finished in June 2009 [68, 58]. For the project, several synthetic population files were created for four different geographical areas. The one for Berlin features 3 320 993 inhabitants that constitute altogether 1 848 897 households and refers to the year 2005, with another scenario designed for 2030. The one of 2005 will be used for the present demonstration example. The file's households are supplied with various attributes such as car ownership and monthly budget as well as a zoning information that tells the area in which the household is located. The distribution is shown in figure 6.2(a) by statistical districts. The actual indication corresponds to the more detailed system of 'traffic subcells', a zoning system maintained by the Statistical Institute of Berlin-Brandenburg (*Amt für Statistik Berlin-Brandenburg*), with an average area of a single subcell of 1 011.3 m^2, 53.8 m^2 being the smallest and 17 272.2 m^2 the largest.

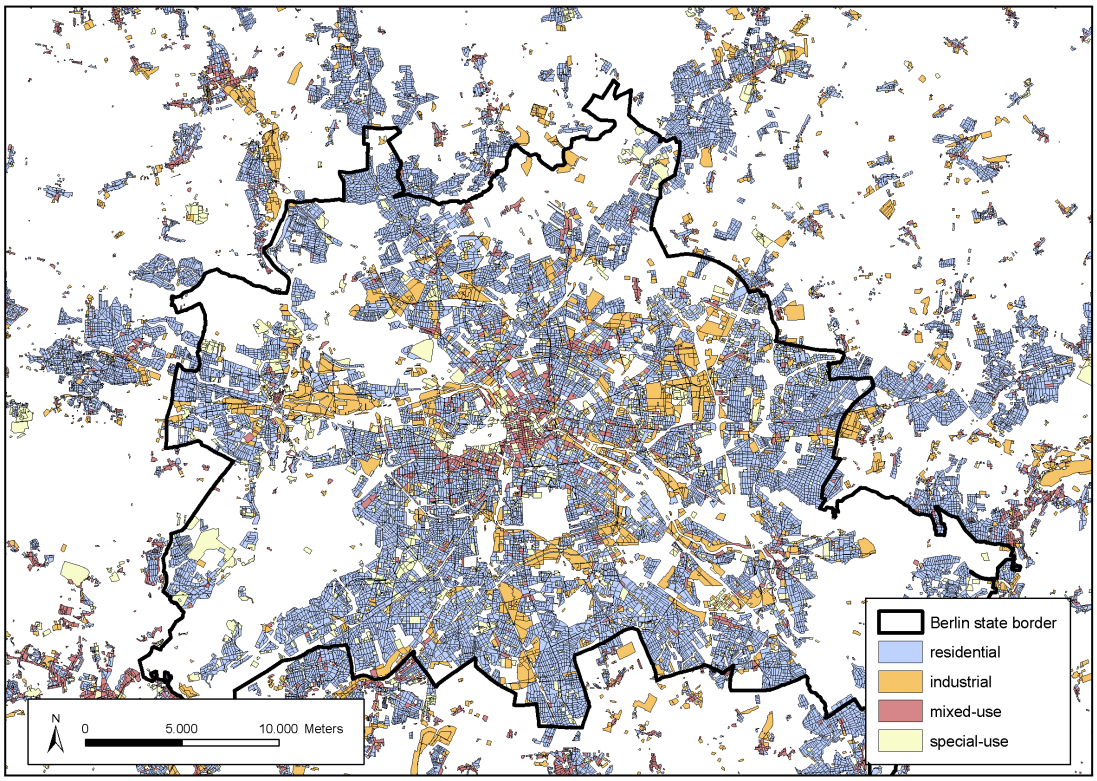
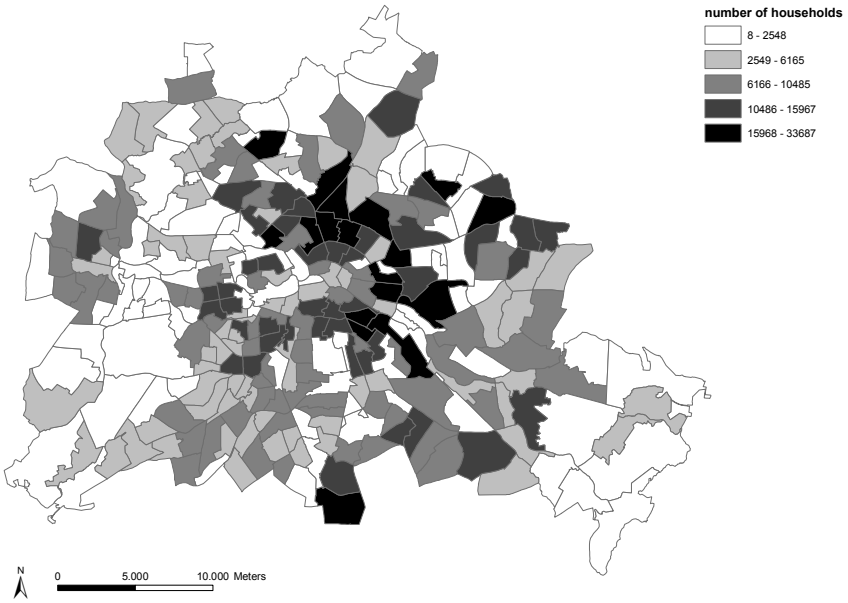


Figure 6.1: Distribution of land-uses in the study area. Source: *Digitales Landschaftsmodel (DLM)* of *Bundesamt für Kartographie und Geodäsie (BKG)*.

The set of buildings is taken from the file of address records that was introduced in section 4.3.2. The disaggregation process results in the assignment of households to buildings, so as every household is specified by street, house number, and postal code. Therefore, an address record is selected for each household based on three assumptions: By definition, a household must first be located within the household's traffic subcell and second, no household should be located outside any of the four land-use types from section 6.2.1. Further, households are more likely situated in residential and mixed areas, only very few, if any at all, should be located in industrial areas and areas of special use. An algorithm is designed to distribute the households accordingly, that is, to distribute households of a cell to buildings that are located inside the land-use areas of that cell. The algorithm takes four ratios as input that specify to which percentage the households of a cell are distributed across the four land-use types of the form "assign 50 % of a cell's households to residential areas within that cell, assign 40 % to areas of mixed use, ...".



(a) Distribution of private households

(b) Distribution of firms

Figure 6.2: Distribution of firms and private households in the study area, by statistical districts. Source: German Aerospace Center (DLR).

The program code that was written for this purpose accepts any target distribution that sums up to 100 %. Rounding errors are handled, guaranteeing a better approximation to the target distribution without supplying or deleting households from the original set. If a land-use type which is supposed to receive households does not exist in a given cell, the algorithm alters the target distribution by assigning these households to a different land-use type. If this type does not exist either, another land-use type is tested, and so on. All four land-use types are checked in successive order (residential → mixed → special → industrial). If all attempts fail, land-use is no longer considered and households will be spread across the whole cell. In this case, households are still assigned to valid address records, ensuring that households are not placed into areas where they would hardly make sense, e. g. inside waterbodies.

For the present simulation, the ratio was defined by estimates. 60 % of the households are placed into residential areas, 39 % into mixed areas, and 1 % into areas of special use. The case that no land-use areas exist in a cell does not occur for the Berlin data, but was implemented to avoid unexpected termination if applying the algorithm to areas for which land-use is not entirely defined.

6.2.3 Economic structure synthesis

For the commercial sector, one can generally notice that data is much more available than for personal data, allowing to choose out of several potential data sources. Firstly, there are commercial data suppliers such as the infas Institute of Applied Social Sciences (www.infas-geodaten.de) and GfK GeoMarketing (www.gfk-geomarketing.de) that maintain firm registers. Their registers usually describe a firm with a number of characteristics such as the firm's postal address, spatial coordinates and further attributes that include the firm's legal form, the firm's number of employees and the firm's economic sector. The main disadvantages of using commercial registers are (1) the high costs that increase significantly with higher spatial resolution and (2) the fact that a methodology for forecasting is missing. Another data source that consists of a listing of individual firms is the national firm register (*Unternehmensregister*) of the Federal Statistical Office [84]. The register contains all firms liable to taxation and keeps information, among others, on the firm's name, its address, legal form, economic sector, number of employees and volume of sales. However, access to it for research purposes is only provided through aggregated statistics.

For the present demonstration example, a synthetic population file is used that combines a commercial data set from the Institute of Applied social sciences (infas)

with several other aggregate data sources [60]. The file that was created as part of the same research project at DLR as the synthetic household population file from section 6.2.2 contains a total of 4.2 million firms for all of Germany, 142 908 of them being located in Berlin, and is designed to be comprehensive. For companies that maintain more than one site, each of them is represented separately, for example if a firm has multiple branch offices. Firms are characterized by their economic sector and the number of employees. A firm's economic sector describes the sector in which the firm's key activity is located and is identified by WZ 2003 [85], a standardized classification scheme (see section 5.2.1). A firm's size refers to its number of employees with liability to social security, classified by seven categories, the smallest ranging from one to nine employees and the largest representing one thousand employees and more. The fact that for some firms, all employees are assigned to the firm's headquarter, causes a spatial distortion for those. A firm's location is specified at the level of official municipalities and for some in more detail at the level of statistical areas (see section 4.3.3), thus reaching a granularity lesser than is the case for the synthetic household population file.

The firm population is disaggregated from the level of zones to individual buildings by applying the same concept that was used prior for the disaggregation of household data. The handling of cases in which the target distribution cannot be met is replaced by a different fall-back procedure: industrial → mixed → special → residential. Firms are also placed into residential areas, following the definition of the DLM's object type catalog that allows small enterprises to be situated in residential areas. While the private household population merely affects the distribution of trips, the number and characteristics of firms directly affect the total traffic volume that will be generated during the simulation. Therefore, the number of firms by economic sector and number of employees is summarized as aggregates in table A.2 on page 121. How the firms are distributed in space is depicted in figure 6.2(b).

6.3 Template logbook extraction

6.3.1 Relevant attributes and case numbers

The template logbook repository from which will be drawn in the present demonstration example is filled with logbooks from the empirical survey *Kraftfahrzeugverkehr in Deutschland* (KiD). Table 6.3 exemplifies a typical logbook record from the survey that was already introduced in section 4.2.1. It shows the movements of a car which belongs to a firm in economic sector D (Manufacturing), with 130 employees and which is based in Frankfurt am Main, Germany. Firms on whose behalf vehicles were operating are characterized by the economic sector (coded in WZ 2003, see table A.1) and the firm size (coded as the number of employees, classified into seven classes, see table 6.4). Further, information on the trip purpose is contained. The coding of trip purposes follows the definition from section 2.1:

1. transportation of goods, materials, machines, ...
2. trip to working location (repair service, delivery and installation service, consulting job, visit, ...)
3. transportation of people (business-related)
4. change in location due to other business-related activity
5. return to company site
6. going to work (going to own company site)

Table 6.3: Example of a logbook record from KiD.

trip	destination type	trip purpose	zipcode	distance	...
-	own company	-	60439	-	...
1	private household	commercial trip (type 2)	60385	33 km	...
2	private household	commercial trip (type 2)	60528	15 km	...
3	private household	commercial trip (type 2)	63303	17 km	...
4	private household	commercial trip (type 2)	60311	12 km	...
7	private household	commercial trip (type 2)	63069	15 km	...
8	other company	commercial trip (type 2)	60433	15 km	...
9	own company	return to company site (type 5)	60439	5 km	...

Source: derived from *Kraftfahrzeugverkehr in Deutschland 2002* (KiD) [100].

7. trip to an educational institution
8. shopping
9. recreation
10. transportation of people (non business-related, private)
11. other private reason
12. return to home

The route on which vehicles were moving can be tracked through postal codes that specify the location of each waypoint and for most, a geo coordinate is provided with higher precision. In addition, waypoints are characterized by their type, for which the survey distinguishes between eight different waypoint types, one of them being 'own company'. In some cases, several waypoints of the latter type exist at different locations. For those, a ninth type is introduced:

1. terminal, station, port, airport
2. forwarder
3. construction site
4. own company
5. external / other company
6. private household (where the service is provided)
7. other business-related destination
8. private destination (e.g. restaurant, home, ...)
9. other site of own company (not originally contained in KiD and introduced here)

Considering the fact that the participants in the study were spread over entire Germany, a question is if all of the logbooks that are contained in KiD can be used for simulating the Berlin area or if only some of them should be considered. If, for example, businesses from Munich show different behavior patterns from those in Berlin, their logbooks will cause misleading results. A study into the effects was carried out

in a joint effort at the Institute of Transport Research (VF) at the German Aerospace Center (DLR) and is currently in the process of publication. The results showed that the type of a region either has low or no influence on the travel pattern, suggesting that vehicle logbooks from across Germany may be used. Notably, the economic sectors where no correlation was found also correspond to those with lowest case numbers and where extrapolations are needed.

6.3.2 Repository setup

For the present demonstration example, the logbook repository is filled with all logbooks from entire Germany, allowing to test the algorithm with a large variety of template logbooks. The case numbers that are obtained are displayed in table 6.4, showing the total number of logbooks in KiD that have one or more commercial passenger transport trips. Cells that are highlighted in light and dark gray indicate that the number of logbooks for the corresponding combination is small, although 0.1 % or more of all firms in the synthetic world belong to the combination. For cells colored in light gray, the number of logbooks is below 100, while for cells colored in dark gray, the number of logbooks is below 50. In the equivalent, for all cells with white background, the number of firms is either below 0.1 % and/or the number of logbooks is greater than 100. And for all combinations with at least 0.1 % of firms, the number of logbooks is greater than 20. These template logbooks are now preprocessed by executing the following successive steps:

1. **Filter according to trip purpose**. The first step is straightforward and takes into consideration the fact that some logbooks do not contain commercial passenger transport trips. A logbook might for example belong to a truck that was just delivering goods during the whole day. As the main interest for the current simulation run is commercial passenger transport, logbooks that only contain non commercial passenger transport trips are not relevant and are therefore removed. Logbooks in which just some trips belong to the category 'commercial passenger transport' are kept to avoid that relevant trips go lost.

2. **Evaluation of the geometrical shape**. In a second step, information on a logbook's geometrical pattern, i.e. the z values as defined in chapter 5.2.2 and in figure 5.4) in particular, is generated. The pattern of a logbook is indirectly provided through the geo coordinates and postal codes that come with each logbook,

Table 6.4: Logbooks with one or more commercial passenger transport trips in KiD by economic sector and firm size (number of employees). Highlighted cells indicate that the number of logbooks for the corresponding combination is small, although 0.1 % or more of all businesses in the synthetic world belong to the combination. For cells colored in light gray, the number of logbooks is below 100, while for cells colored in dark gray, the number of logbooks is below 50. In the equivalent, for all cells with white background, the number of firms is either below 0.1 % and/or the number of logbooks is greater than 100.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1-9	208	3	12	465	141	1948	655	28	153	64	444	129	71	104	260
10-49	124	1	23	391	210	1823	469	31	227	30	307	311	28	162	205
50-99	22		5	97	73	384	123	7	43	13	75	88	6	35	65
100-249	12		13	112	97	222	75	4	56	8	74	67	6	48	62
250-499	1		1	57	57	39	29	2	21	7	29	43	1	21	20
500-1000	1		1	51	38	30	23	1	14	10	17	13		11	10
> 1000			5	72	30	10	15		21	9	14	14		5	2
n.s.	17		7	81	93	234	100	4	56	14	64	84	10	37	62
total	385	4	67	1326	739	4690	1489	77	591	155	1024	749	122	423	686

dark gray: less than 50 logbooks with the economic sector/business size combination exist, although 0.1 % or more firms in the synthetic world report this attribution.

light gray: less than 100 logbooks with the economic sector/business size combination exist, although 0.1 % or more firms in the synthetic world report this attribution.

Source: derived from *Kraftfahrzeugverkehr in Deutschland 2002* (KiD) [100]. See table A.1 on page 115 for a description of all economic sectors.

because by lining up all of the waypoints, a line can be drawn along which the vehicle was moving.

The z values were thus calculated based on the geo coordinate pairs of the KiD attributes. If these attributes were found empty, a coordinate was generated pointing to the geometric center of the waypoint's postal code district. Some cases provide information on the KGS district, but not on the postal code. Here, a coordinate was generated that points to the municipality's center. Coordinates are generated only if a validity check of postal code and KGS district was successful. In very few cases, a trip's geo coordinates did not correspond to its other position information and thus its coordinates were re-generated based on postal code or KGS data.

Figure 6.3 visualizes a logbook and names the attributes from which the location information was extracted.

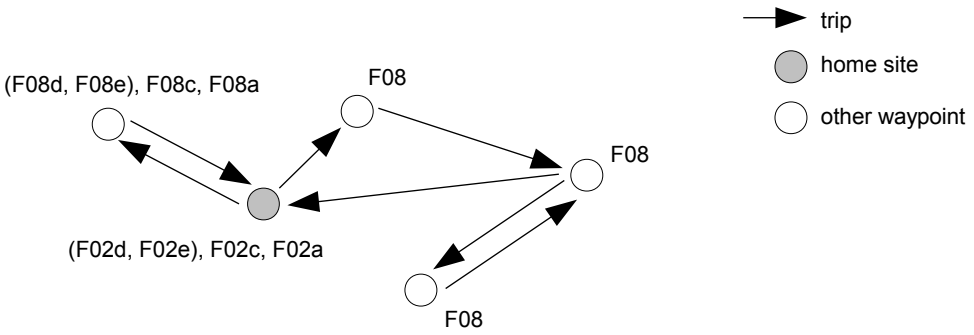


Figure 6.3: Waypoint coordinates and corresponding KiD attributes. The figure shows a typical logbook pattern from the empirical survey *Kraftfahrzeugverkehr in Deutschland 2002* (KiD) [100]. The labels that are displayed next to each waypoint tell the original KiD naming of the attributes containing the positioning information.

3. **Resolving branch office vs. own business**. A third step considers that KiD does not distinguish between a firms headquarter and various branch offices, which is a possible explanation why in some cases several waypoints are marked as 'home business' although their geographical position differs [53]. The issue is resolved by introducing an additional waypoint type 'branch office'. If a logbook contains waypoints of type 'home business' with different geo-locations, the coordinate pair that appears most often keeps the original attribution, all others receive the attribution 'branch office'. In case that there is no coordinate that occurs more often than others, the first is considered to be the home business, all following ones become 'branch office'.

4. **Trip description in varying levels of detail**. A drawback is that for trip twelve and upwards, participants were no longer requested to describe their trips in full detail. Since no information on the destination's type is given for such, at the most eleven trips become part of the simulation. Figure 6.4 shows that the majority of logbooks with one or more commercial passenger transport trips show small trip rates and that only 5.3 % were reported to have twelve or more trips. An alternative to provide logbooks in full length for simulation is by synthetically generating data in a way that is consistent with the known total number of trips and the known total number of kilometers driven that day.

   However, also trips before the twelfth trip must be adapted, as they have not

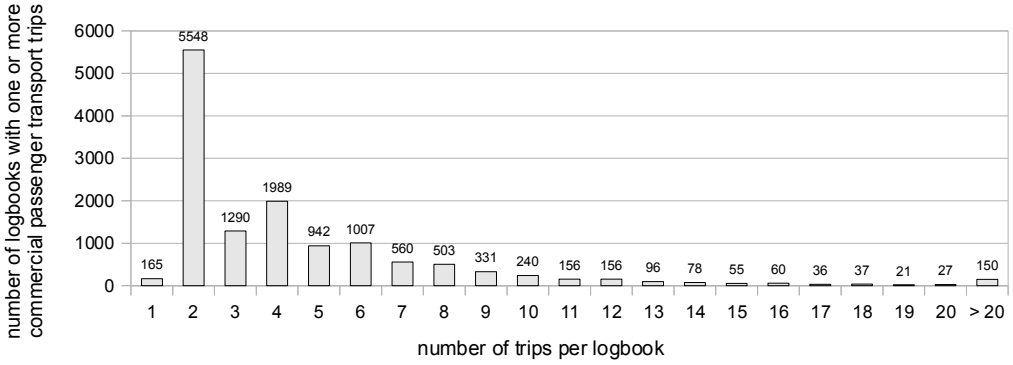


Figure 6.4: Number of trips per logbook (attribute 'K31') for logbooks with one or more commercial passenger transport trips (type 2 and 3). Logbooks that have twelve or more trips represent altogether 5.3 % of all logbooks. Source: *Kraftfahrzeugverkehr in Deutschland 2002* (KiD) [100].

been reported properly in some cases. Occasionally, participants of the study summed up several trips into one trip, causing some logbooks contain records that actually represent multiple trips. In total, 949 out of 12 705 logbooks that remain from the previous steps are marked to contain 'multiple trips' that are combined into one. Since no coordinates are provided for the missing waypoints and thus the actual shape is not provided for such, their spatial pattern cannot be reconstructed. Therefore, such logbooks are filtered out.

5. **Missing values.** For some trips, although only in a very few cases for the logbooks that remain, attributes such as a destination's type is coded as 'unknown/not specified'. Those records are passed on unchanged into the logbook repository, allowing the simulation algorithm to choose out of all destinations that lay at the given distance. If it is a trip's purpose that is unknown, the trip is generated equivalently as a 'trip of unknown purpose' during simulation.

Trips with origin and destination of type 'home site' are kept. In effect, the simulation output will contain trips that start at the home site and immediately return to the starting position. All steps were implemented as a PL/SQL script that is attached in listing A.1.

6.4 Mobility ratio extraction

The mobility ratios are extracted from the representative survey of the German service industry that was introduced earlier in section 4.2.2. With respect to the underlying answer set, the extraction is simplified by not distinguishing between multiple mobility ratios. Therefore, one single ratio is determined that represents the number of mobile vehicles of a firm during an average work day.

The ratio is determined in a number of consecutive steps. To begin with, for each of the 27 separate service types that are displayed in table 6.5, the average number of times a service is provided by a company per customer per year is extracted. In addition, participants were requested to specify for each service type where this service is provided. For example, participants could specify whether the service is provided 'at the customer's site', 'at other site' or 'in-house'. Whether services are provided via paper letters, electronic mail or internet, and thus no movement takes place, is on record as well. The two location types that are of interest – as they are the only ones that generate trips – are 'external at customer' and 'at other site'. All others are filtered out.

Participants were also requested to specify their number of customers. As before, this is provided for each of the 27 service types. Multiplying both results in the number of times that each service type is provided externally. Assuming that no more than one service is provided to a customer at a time, all service types can be aggregated to determine the number of times services of any type are provided. Under the premise of a second assumption considering that no information about the duration of a service is given, this is equal to the number of times that external sites are visited by firms. That is, it is assumed that customers are visited only once per service. While this is not problematic for services that are not very time-consuming, it is for services that are provided over long periods of time, as they almost certainly require more than one visit. Common examples for the first are repair and installation services, typically completed within hours without requiring additional trips. An example for the latter are financial auditing services that are typically provided during several days or weeks, for which large customers such as banks and investment funds often provide office space for the service provider at the customer's site.

Information on the mode choice of businesses is also part of the survey and is given as percentages of all of a company's commercial trips, requiring a third assumption that the mode choice for services-related trips is equal to the mode choice of all of a

Table 6.5: List of service groups from the empirical survey *Dienstleistungsverkehr in industriellen Wertschöpfungsprozessen*

1	Research and development	(Forschung und Entwicklung)
2	Projection	(Projektierung)
3	Assembly	(Montage)
4	Maintenance	(Instandhaltung: Produktionsmittel/Maschinen)
5	Software development	(Softwareentwicklung)
6	Market research	(Marketing)
7	Advertisement	(Werbung)
8	Procurement	(Einkauf/Beschaffung)
9	Sales and distribution	(Vertrieb)
10	Consignment	(Kommissionierung)
11	Warehousing	(Lagerhaltung)
12	Shipping	(Versand)
13	Customer training	(Kundenschulung)
14	Cleaning	(Reinigung)
15	Security service	(Sicherheitsdienst/Werkschutz)
16	Cafeteria	(Kantine)
17	IT	(Datenverarbeitung/IT)
18	Accounting	(Rechnungswesen/Buchhaltung)
19	Legal advice	(Rechtsberatung)
20	Assurance	(Versicherungen)
21	Financial service	(Finanzdienstleistung)
22	Consulting	(Unternehmensberatung)
23	Financial auditing	(Wirtschaftsprüfung/Steuer)
24	Human resource management	(Personalwesen)
25	Further education	(Weiterbildung/Mitarbeiterqualifizierung)
26	Facility management	(Gebäudemanagement)
27	Waste disposal	(Abfallentsorgung)

Source: Derived from *Dienstleistungsverkehr in industriellen Wertschöpfungsprozessen* [43].

company's commercial trips. From the modes that were surveyed, the ones of interest are trips with motorized vehicles due to the focus on road traffic in the present simulation example. In the survey, these trips are marked as 'trips with vehicles owned by the company', 'trips with private vehicles' and 'trips with rental cars'. As a result, the number of visits to external sites with motorized vehicles per company is obtained. Since this is annually, it is next divided by the average number of working days per year in order to receive daily numbers. All calculations are summarized in the formula

$$x = \frac{\left(\sum_{d \in D} (a_d + b_d)\, c_d\right) (m_1 + m_2 + m_3)}{252} \tag{6.1}$$

where x is the average number of visits to external sites per day with motorized vehicles,

D is the set of 27 service types listed in the survey, a_d is the number of times a service $d \in D$ is provided at a customer's site per year, b_d is the average number of times a service d is provided externally 'at other site' and c_d is the average number of customers for service d per year. m_1, m_2 and m_3 are the percentages for the three road modes 'company-owned vehicle', 'private vehicle' and 'rental car'. The constant 252 is chosen as the average number of working days per year. After performing a plausibility analysis, the results that are displayed in table A.3 are obtained.

Next, the empirical survey KiD is used to determine the number of customers that are visited per vehicle during an average work day. For this purpose, the number of relevant trips in each logbook is counted. A trip is considered relevant if the destination's type and the trip's purpose meet the following two criteria. First, a trip's purpose must be either of type

- transportation of goods, materials, machines, ...,
- trip to working location (repair service, delivery and installation service, consulting job, visit, ...),
- transportation of people (business-related), or
- change in location due to other business-related activity

and second, the trip's destination is not of type

- own company or
- private destination (e. g. restaurant, home, ...).

The results displayed in table A.4 represent the average number of visits to customers of logbooks with one or more commercial passenger transport trips, by economic sector and firm size. By dividing a firm's average number of visits to external sites (table A.3) by the average number of visits to customer sites (table A.4) of vehicles with the firm's economic sector and size, the number of vehicles that operate per day in order to undertake that many customer visits is obtained. The final result is displayed in table 6.6.

For most economic sectors that are displayed in table 6.6, the classification's granularity for the attribute firm size lacks sufficient detail. For example, the majority of vehicles of economic sector I (transport, storage and communication) operate on behalf of firms with 100 to 249 employees and only few vehicles operate on behalf of

Table 6.6: Number of vehicles operating per day per business, by economic sector and number of employees, rounded up to full integers.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1 - 9 employees	0	0	0	1	1	1	1	0	2	2	1	0	3	0	1
10 - 49 employees	0	1	0	1	1	1	4	0	4	4	5	0	1	1	1
50 - 99 employees	0	0	1	1	1	13	43	0	1	6	12	0	2	0	25
100 - 249 employees	0	0	0	38	1	9	1	0	132	76	15	0	0	4	11
250 - 499 employees	0	0	1	1	9	150	15	5	1	12	16	88	0	0	5
500 - 1 000 empl.	0	0	5	1	0	0	1	0	0	6	70	0	0	1	4
$>$ 1 000 employees	0	0	1	0	0	0	0	0	1	292	12	1	0	0	82

Source: Derived with the presented procedure by refering to the two empirical surveys *Dienstleistungsverkehr in industriellen Wertschöpfungsprozessen* [43] and *Kraftfahrzeugverkehr in Deutschland 2002* (KiD) [100]. See table A.1 on page 115 for a description of all economic sectors.

firms of other classes. The same applies for most other economic sectors. Only in sector K (real estate, renting and business activities), the values are distributed over all classes. Despite these downsides, the classification scheme is chosen to correspond to the one of the synthetic economy from section 6.2.3 with the advantage of having the same scheme consistently applied through all modeling steps. Notably, the two empirical surveys *Kraftfahrzeugverkehr in Deutschland* (KiD) and *Dienstleistungsverkehr in industriellen Wertschöpfungsprozessen* in principle allow to classify with more detail.

6.5 Simulation and result analysis

6.5.1 Parameter configuration

Starting with the waypoint type constraint c_1, the simulation parameters are set as follows:

1. If the trip's destination is of type 1 (terminal, station, port, or airport), the search is limited to address records located within areas of land-use type 'special use'.
2. If the trip's destination is of type 2 (forwarder), businesses of economic sector 'I' (transport, storage and communication) are considered.
3. If the trip's destination is of type 3 (construction site), any address record can be selected.
4. If the trip's destination is of type 4 (home site), the trip must be directed to the business on whose behalf the vehicle operates.
5. If the trip's destination is of type 5 (external company), any business can be selected except the vehicle's own.
6. If the trip's destination is of type 6 (private household), the trip is directed to a household.
7. If the trip's destination is of type 7 (other business-related destination), any address record is considered, since no information is provided in KiD that allows to further specify the location type.
8. The same applies to destinations of type 8 (private destination), which are interpreted similarly. Any address record can be selected.
9. If the trip's destination is of type 9 (branch office), the trip must lead to a business whose economic sector is that of the firm to which the vehicle belongs, with the idea in mind that although the synthetic firm file does not contain relations on how branch offices relate to each other, it is assumed that both firms must be listed with the same economic sector.

The items are formalized in table 6.7. The mapping's granularity is subject to the available data. For example, an alternative for trips with destination of type two

Table 6.7: Constraint specification for location assignment (c_1)

id	location type	restriction	C^{c_1}
1	terminal, station, port, airport	any location within land-use 'special'	$\{c \in C \mid lu(c) = \text{'special'}\}$
2	forwarder	any firm with economic sector 'I' (transport)	$\{b \in B \mid es(b) = \text{'I'}\}$
3	construction site	any location	C
4	own company	own company exclusively	$\{b \in B \mid b = x\}$
5	other company	any firm except the own	$B \setminus \{x\}$
6	household	any private household	H
7	other business-related destination	any location	C
8	private destination	any location	C
9	branch office	any firm sharing own economic sector	$\{b \in B \mid es(b) = es(x)\}$

C is the set of locations.
B is the set of businesses.
H is the set of private households.
x is the business to be processed.

(forwarder) would be to have them point to firms whose economic sector corresponds to 'cargo handling and storage' (NACE subsection 63.1). Since the virtual world does not support this level of detail, the upper next level (I: 'transport, storage and communication') is chosen.

The distance constraint c_2 is set to consider the direct airline distance between the origin and destination for the current simulation run. Epsilon is set to ± 50 meters, specifying that the length of a generated trip can differ from the length of the original template logbook within this range. Epsilon is not expressed as a function of the trip length with the intention to test the algorithm with small epsilons. The algorithmic implementation assumes that half of ϵ is sufficient in most cases, and only if no elements are found (no elements exist that meet the constraint c_1 within that corridor) or subsequent trips fail, the search corridor is expanded to the full epsilon range. Larger values for epsilon would lead to higher rates of applicable template logbooks.

The sorting criterion z will sort the remaining waypoints according to how well they reflect the original trip's distance-to-homesite in the template logbook and waypoints are allowed to differ from the original distance-to-homesite without limit. The mapping between firms and template logbooks is chosen to resemble equal attribution defined in formula 5.32 on page 48. Multiple mobility ratios per firm are not differentiated.

Template logbooks are selected from the set of potential logbooks by random and no weight function is applied. A 100 % sample, that is, all firms located in the study area, will be processed.

6.5.2 Experimental results

The simulation results will be analyzed by first investigating into the formal success rate, then analyzing aggregates and finally showing examples of generated trip chains on individual level.

Formal analysis

In total, 520 637 trips were generated during the simulation and 265 367 of these were trips of type commercial passenger transport. The trips were made by 133 538 vehicles on behalf of 97 936 firms. For each vehicle, a template logbook was selected. In 97.7 % of all cases, the template logbook could be successfully applied to the synthetic world and in the remaining 2.3 %, the assignment failed. For these cases, the constraint definition in conjunction with the characteristics of the synthetic world did not allow to apply the selected template logbook. In another 25 492 cases, the template logbook that was chosen could not be assigned with respect to the fact that it did not contain waypoints of type homesite. By this, no pivot point is defined that specifies where to place the template logbook. For these, one of the measures that were proposed in section 5.5.1 need to be applied: (1) random dispersion, (2) employee home/work relationship modeling or (3) impact analysis. As anticipated in section 5.5.1, a significant share of the logbooks in question were operating far from their home location. Many trips were therefore made outside the study area and hence are not relevant, which is why no further steps are taken here to consider them. If alternative pivot points were to be supplied for those in question, their application can be expected not to differ from the results that were observed for the rest of logbooks. The formal success rates are summarized in table 6.8.

Aggregate analysis

The generated traffic volume is expressed in absolute numbers in table 6.9. It shows the total number of trips that were generated during the simulation per vehicle type. It must be noted that the actual numbers that were obtained should not be considered to be precise estimates since they reflect several aspects that are specific to the input

Table 6.8: Formal analysis of the simulation results. The assignment failed if a template logbook or one of its sub-logbooks could not be assigned.

Log event description	absolutes	percentages
Template logbooks successfully applied	130 407	97.7 %
Logbooks that did not fit to the synthetic world	3 131	2.3 %
a) Closed trip chain could not get assigned.	1828*	
b) Open trip chain could not get assigned.	1576*	
Total number of vehicles processed	133 538	100 %
Total number of firms processed	97 936	
Logbooks with no stop at home business (not implemented)	25 492	

* a) and b), when taken together, exceed the number of 3 131 failed template logbooks, because for logbooks with several sub-logbooks, each that fails is counted separate.

data. The results are not empirically sound because of the limited data that is available for calibration and the relatively small effort that was spent on data collection and preparation. For comparison, Rossi et al. interviewed twenty project managers in the United States responsible for the development and data gathering process of regional activity-based traffic models and concluded that the model development usually takes between two to three years and costs US $600 000 to $800 000 in additional consulting fees [79].

Table 6.9: Number of simulated trips, by vehicle type

Vehicle type	number of trips	thereof trips with commercial purpose (KiD trip purpose 1 to 5 and 21)
motorbikes	10 212 (2 %)	9 677
automobiles	311 044 (59.7 %)	246 703
trucks up to 3,5 t of payload	191 800 (36.8 %)	181 126
trucks with more than 3,5 t of payload	2 854 (0.5 %)	2 854
semitrailer trucks	457 (0.1 %)	457
overland buses	3 052 (0.6 %)	3 052
emergency vehicles	88 (0 %)	88
others	1 130 (0.2 %)	1 102
total	520 637 (100 %)	445 059

However, despite these limitations, the shares per vehicle type in the simulation remarkably correspond to the ones published by Steinmeyer and Wagner in [89] and [90]. According to their studies, the number of passenger car trips in Berlin was estimated to account for 62 % of total traffic, and for trucks with up to 3.5 tonnes of payload the number of 34 % is given. In comparison, passenger car trips account for 60 % of the trips in the current simulation run and trucks with up to 3.5 tonnes of payload accounted for 37 % in the simulation (see figure 6.5). A significant difference between the simulation's distribution of vehicle types and that of empirical figures would hint that the grouping of firms by size and economic sector did not result in perfectly homogeneous behavior pattern sets, suggesting to distinguish between multiple agent types per firm and/or to apply a weighting factor for taking into account the discrepancy.

A further comparison of the two estimations shows that neither the absolute numbers differ. So was the number of trips per day estimated by Steinmeyer and Wagner to be roughly 510 000 trips, with a range of tolerance between 410 to 610 thousand trips [89]. As shown in table 6.9, the sum of all trips in the current simulation lies in this range and accounts for 521 000 daily trips.

In order to analyze the number of vehicle kilometers driven, a sample of 7 170 trips from those that were generated by the simulation was drawn. A route within the road network was then generated for each by querying the shortest travel time algorithm

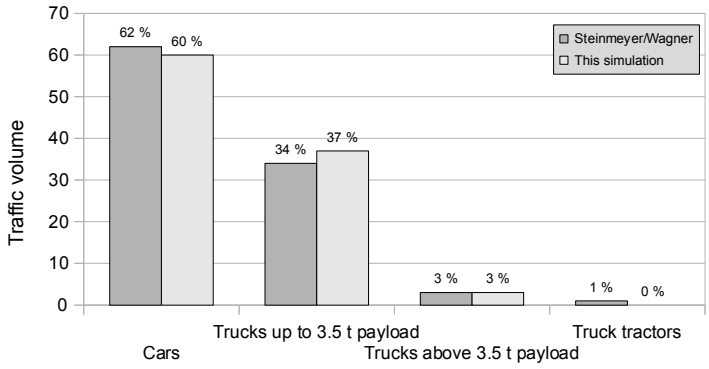


Figure 6.5: Traffic volume per vehicle type in percent. Comparison between estimates by Steinmeyer and Wagner [89, 90] and simulation results.

of the XML web service Google Maps [39]. The results are displayed in table 6.10, showing that the total number of vehicle kilometers travelled make up approximately 13.5 million kilometers by this projection. As such, the simulation also narrowly corresponds to the estimate by Steinmeyer and Wagner of 11 million vehicle kilometers with a range of tolerance between 9 and 13.5 million. A comparison by vehicle type shows that the share of kilometers travelled with automobiles represents about 70 % in both studies (70 % in this simulation and 71 % in Steinmeyer and Wagner). Only for trucks, the shares differ slightly with 28 % of trucks with up to 3.5 tons of payload in the present simulation opposed to 18 % respectively. Considering the high degree of similarity in most measures, the present simulation is not further calibrated.

The numbers by Steinmeyer and Wagner are chosen for verification because they are comparable in that the two study areas are identical and only traffic on behalf of firms that are located inside it is considered. Incoming traffic from companies from outside is not included. The two calculations are further identical in that they both consider private and non-private trips of these firms and also, both cases generate the transport demand for an average work day (a regular Tuesday, Wednesday, or Thursday). Further, their estimates are solid in that their methodology successfully passed the scientific peer-review process twice [89, 90], including a presentation at the Transportation Research Board Annual Meeting in 2006. A drawback is, however, that their numbers are based on several assumptions that limit their suitability for comparison. The main difference is that the results by Steinmeyer and Wagner

Table 6.10: Routed vehicle kilometers travelled, by vehicle type and for all vehicles. Projection based on a sample with n = 7 170, no projection made for small case numbers.

Vehicle type	N	n	avg. routed distance per trip in km	projected vehicle kilometers	
				absolute	percentages
motorbikes	10 212	123	13.18	134 573	1.0 %
automobiles	311 044	4 196	30.61	9 521 326	70.3 %
trucks up to 3.5 t of payload	191 800	2 756	19.57	3 753 691	27.7 %
trucks with more than 3.5 t of payload	2 854	37	-	-	-
semitrailer trucks	457	3	-	-	-
overland buses	3 052	38	-	-	-
emergency vehicles	88	3	-	-	-
others	1 130	14	-	-	-
all	520 637	7 170	26.0	13 536 562	100.0 %

only consider trips on behalf of commercially registered vehicles, hence commercial trips that are made with vehicles registered to private holders are not covered. On the contrary, the present simulation did not differentiate between vehicles registered to private and commercial holders. Hence, the number of trips in the simulation should exceed the ones by Steinmeyer and Wagner with those made by vehicles registered to private holders. The fact that this is not the case might be because the empirical survey *Dienstleistungsverkehr in industriellen Wertschöpfungsprozessen* that was used for the derivation of the mobility ratios emphasized business to business (B2B) relationships, while business to consumer (B2C) relations are not entirely covered. Therefore, a question left unanswered here is whether the amount of missing trips in each approach is equivalent. While the simulation results suggest that they are, an empirical analysis is needed for full validation. Another critical issue is that one of the input data sources that were used by Steinmeyer and Wagner is KiD, representing an overlap of input data between the two procedures. However, their usage of KiD is different in that they referred to KiD for the extraction of aggregates, while the present simulation refers to KiD's raw trip chain records. The rest of input data sources do not overlap. So did Steinmeyer and Wagner use the firm register of the Statistical Institute of Berlin and the vehicle registration numbers of the Federal Motor Transport Authority for their approach. On the contrary, for the simulation, a firm register obtained from infas Geodaten (see section 6.2.3) and trip rates from the empirical survey *Dienstleistungsverkehr* (see section 6.4) were used.

The methodology definition in chapter 5 specifies what variables ultimately affect the simulation's output. As stated there, the two key figures 'number of trips' and the 'kilometers driven' are dependent from three input variables:

1. The set of firms B. A greater amount of firms will lead to a higher number of trips. Also, the set's composition with economic sector and firm size determines the selection of template logbooks, in consequence affecting the trip volume and the kilometers traveled.

2. Further, the mobility ratios $m(b)$ or $m_i(b)$ respectively, since the number of vehicles that operate on behalf of firms directly affects the two key figures.

3. And finally, the characteristics of the template logbooks contained in the repository affect the output by their number of trips $|T_l|$ and their distance covered $d(l)$ of a logbook l.

The parameter ϵ can be assumed not to affect the two key figures significantly despite of the actual size that is chosen, as with high case numbers, trips with large (positive) epsilons are compensating for smaller (negative) ones. Neither are spatial data sources listed above despite their importance to the present methodology, because they only affect the distribution of trips without changing their distance, thus showing no effect on the trip volume nor the kilometers driven.

Of the variables with impact, the input data used by Steinmeyer and Wagner overlap with the third and do not intersect with the first and the second. On the one hand, this shows that the comparison data and the input data are not fully disjunct, hence not fulfilling the model theoretic requirements that were given in section 4.1. The limited data available makes a compromise necessary, for the cost of impeding full validation. On the other hand, the comparison also shows that despite the input data and its usage not being fully identical, the two approaches lead to similar results, suggesting that the two methodologies correctly reproduce the input data and that the simulation is free of programming bugs and incorrect methodology design.

Another essential aspect in output accuracy is the spatial distribution of trips. This is contrary to the methodology by Steinmeyer and Wagner, which does not specify how the trips distribute. The high formal success rate that was obtained in the simulation with roughly 98 % successfully mapped logbooks raises the question whether the set of constraints was set too weak, giving the algorithm far too many choices that might have caused a pointless dispersion of trips in space. Hence, the spatial distribution of trips is analyzed next in figure 6.6. The figure displays the number of incoming and leaving trips for every traffic cell. As shown, the majority of trips are generated in the city districts and the outer suburbs show lesser degrees of incoming and outgoing trips. Thereof exceptional are some districts with higher than average trip rates that lie at the study area's outside boundary. This effect is plausible because in the present simulation, the study area captured the entire synthetic world, causing a distortion for trips whose distance barely fit into it. Hence, such trips were more likely placed into the few outer districts that were defined, overestimating the actual trip volume for these districts. The effect can be avoided by having the study area be a smaller part of the synthetic world as suggested in section 5.2.1.

The departure time histogram in figure 6.7 shows the distribution of trips over twenty-four hours for all trips and in figure 6.8, the trips are split by trip purpose. Since no equivalent data from other research is available, the curves cannot be compared with other data. Nonetheless it can be said that the histograms generally follow the typical

shaping with morning and afternoon peaks, showing that the program code apparently worked well in terms of correctly reproducing the input template logbooks. Since the code "clones" these templates, any major difference such as no trips during the afternoon would hint a possible programming bug, while any minor difference such as a shift from too many trips in the morning hours to too few in the evening can be caused by the incorrect selection of template logbooks, e. g. if a specific non-representative logbook were chosen disproportionately frequently.

More analyses can be made with the output data. So does figure 6.9 line up all trips on the x-axis in the order of their distance, showing that very few trips show long distances and the majority of trips are of short distance with less than 100 kilometers. The average number of trips per trip chain in the simulation correspond to 3.69 trips. By this, the rate is found to be within the corridor of 3.6 to 3.8 trips per day for weekdays that was observed in the recent empirical survey *Mobilität in Deutschland 2008* (MiD) [35]. However, a full validation must be subject to further research. The analyses that were given so far have been primarily qualitative. An

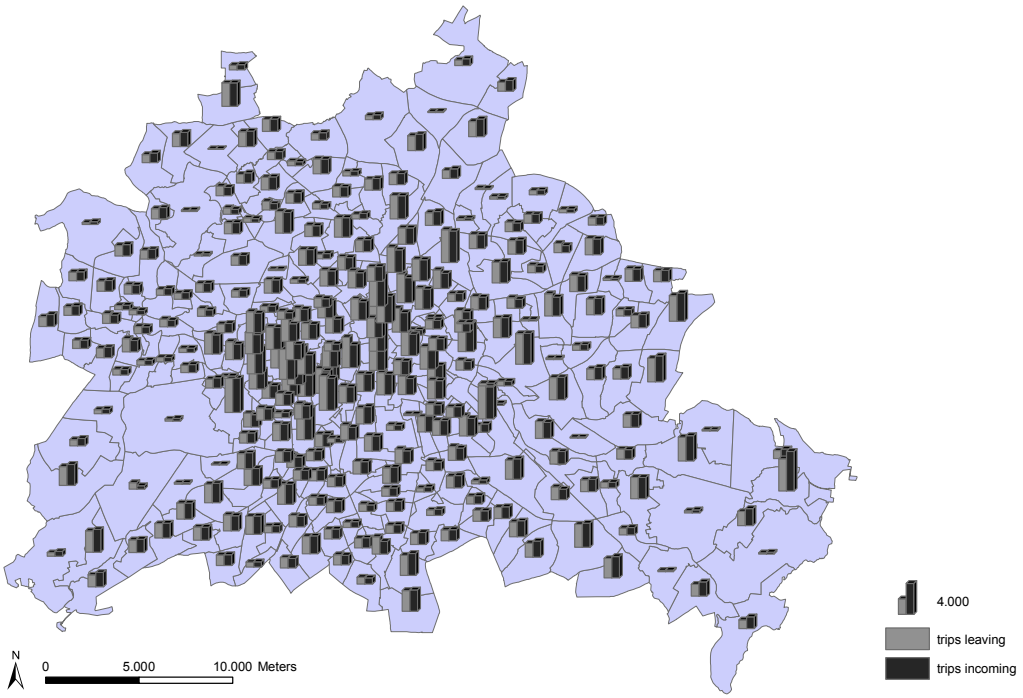


Figure 6.6: Spatial distribution of incoming and leaving trips by statistical area.

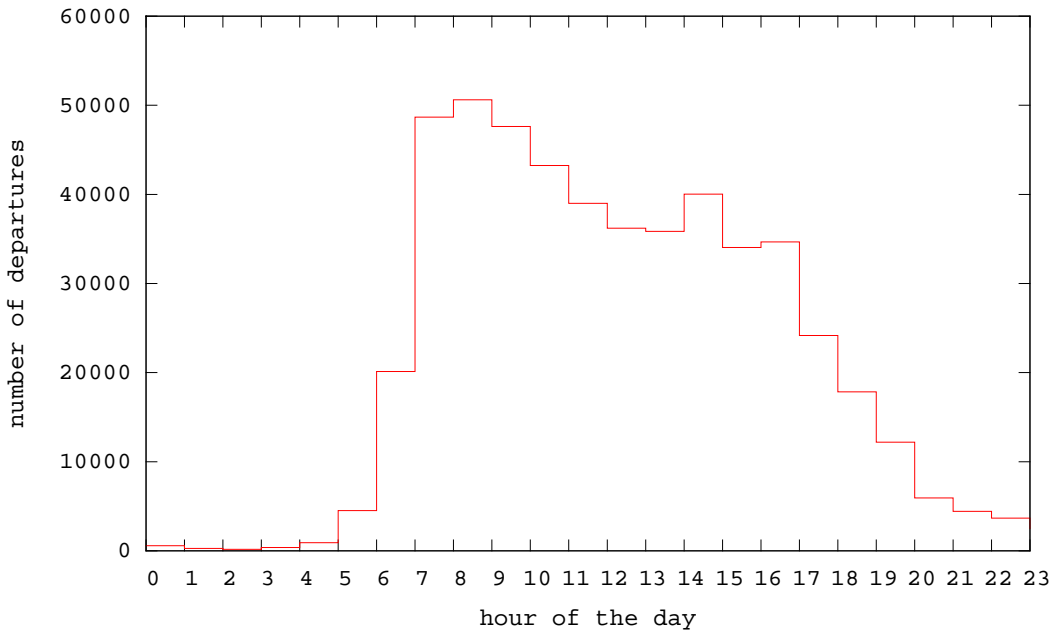
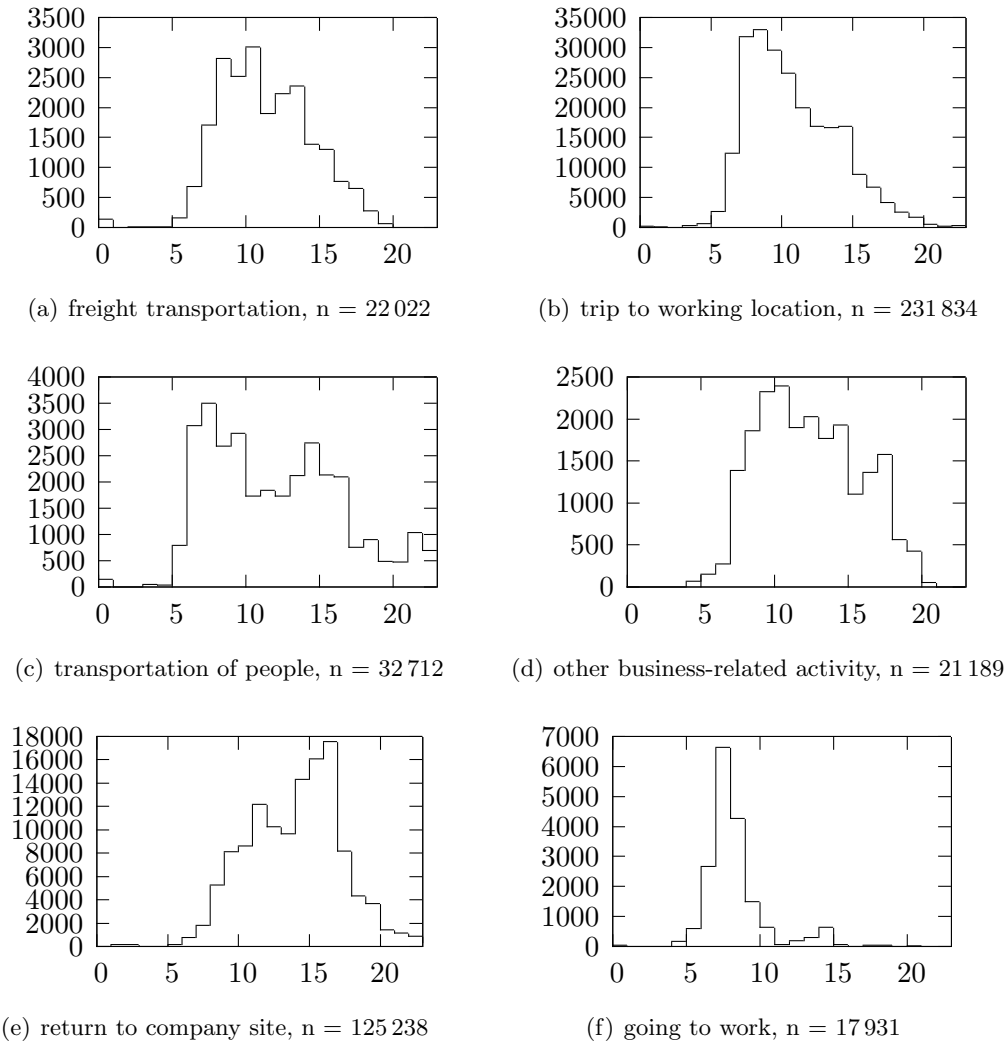


Figure 6.7: Departure time histogram, n = 507 662

ultimate model verification requires to link the methodology to road traffic simulators for direct comparison with life traffic counting data from the road network.

Individual trip chain analysis

A notable feature of the present methodology is the generation of individual trip chain records, such as the one depicted in listing 6.1. The listing shows the trip chain record in Matsim XML, the file format previously introduced in section 3.2.3. The trip chain consists of five individual trips (or legs), the first starting at 05:40 am and arriving at the last waypoint at 15:55 pm. Note that the empirical survey KiD, which provided the template logbooks that were used for the simulation, only specifies the trips of a vehicle without giving much information on the activities between the trips. In consequence, the activity types are coded as 'not specified'. The coordinates can be used to transform the dayplan into a route. Figure 6.10 shows a route that was generated using the 'shortest route' mode of the web service Google Maps.



(a) freight transportation, n = 22 022

(b) trip to working location, n = 231 834

(c) transportation of people, n = 32 712

(d) other business-related activity, n = 21 189

(e) return to company site, n = 125 238

(f) going to work, n = 17 931

Figure 6.8: Departure time histograms by trip purpose.

6.5.3 Computational speed

Computation of the 100 % sample accounted for 42 hours, the equivalent of 1.13 seconds per firm and template logbook and 0.29 seconds per trip on an Intel Xeon 3.00 GHz machine with 12 GB of RAM running Ubuntu Linux 4.3.2 64-bit and PostgreSQL 8.3.7. While this is almost two days, the run time seems acceptable because during the code development, the primary focus was on the actual program logic. Not much of the potential for speed improvements has been exploited because of several reasons. One was that concentrating on the model logic allowed to keep the program code simple

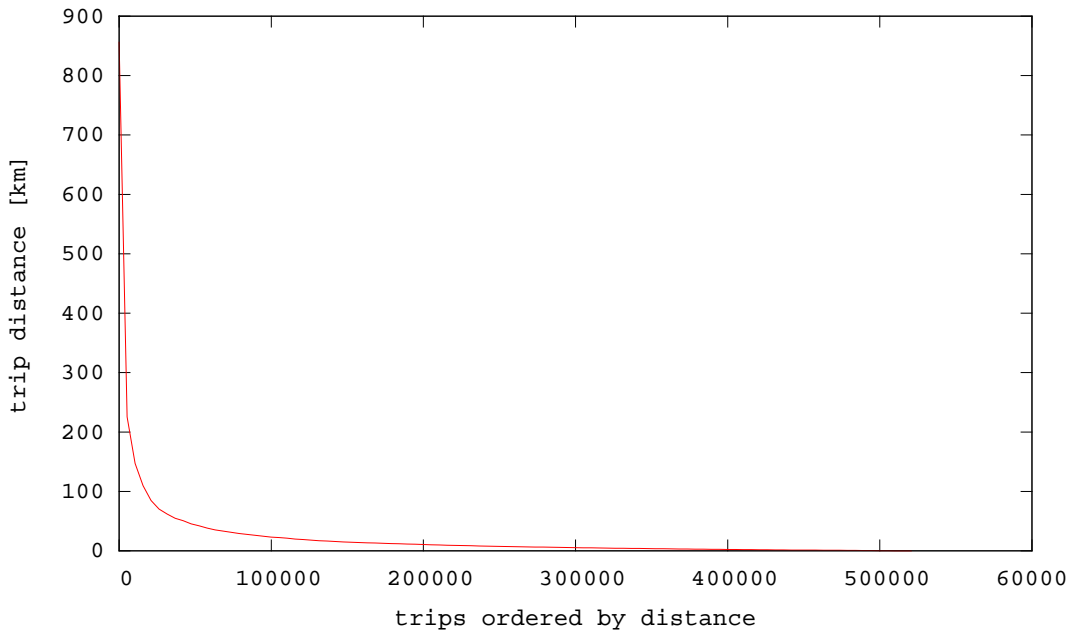
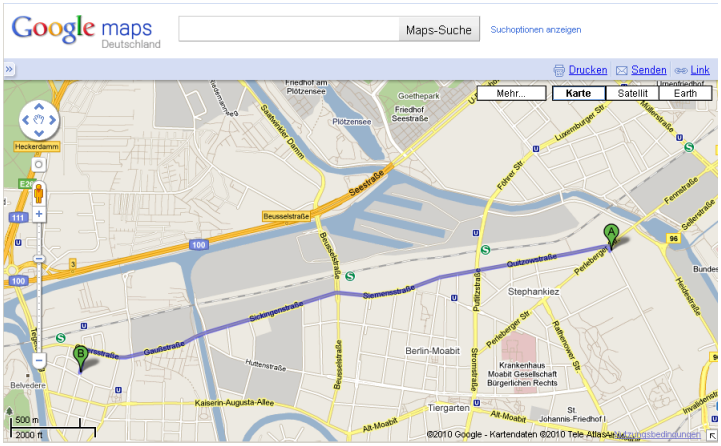


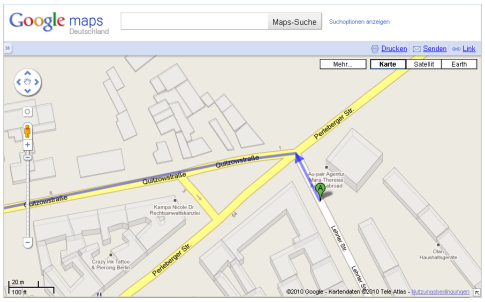
Figure 6.9: Trips by distance, n = 520 637

Listing 6.1: Example of a generated trip chain record in Matsim XML

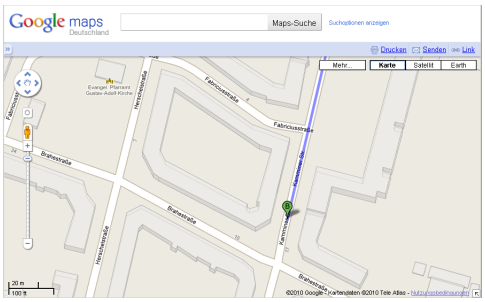
```
<plans name="plans file" xml:lang="de-DE">
  <person id="1">
    <plan selected="yes">
      <act type="ns" x="13.358057" y="52.535817"
             start_time="00:00:00" end_time="05:40:00" />
      <leg mode="car" dep_time="05:40:00" arr_time="06:32:00" />
      <act type="ns" x="13.074649" y="52.962143" start_time="06:32:00" />
      <leg mode="car" dep_time="09:10:00" arr_time="09:30:00" />
      <act type="ns" x="13.358057" y="52.535817" start_time="09:30:00" />
      <leg mode="car" dep_time="11:00:00" arr_time="11:15:00" />
      <act type="ns" x="13.301205" y="52.527727" start_time="11:15:00" />
      <leg mode="car" dep_time="13:00:00" arr_time="13:20:00" />
      <act type="ns" x="13.203236" y="52.519690" start_time="13:20:00" />
      <leg mode="car" dep_time="15:40:00" arr_time="15:55:00" />
      <act type="ns" x="13.303030" y="52.551856" start_time="15:55:00" />
    </plan>
  </person>
</plans>
```



(a) Complete route.



(b) Origin



(c) Destination

Figure 6.10: Route example showing the fourth trip of the trip chain from figure 6.1 using the Google Maps shortest path search [39].

and easy to maintain. More importantly, the fact that the main loop iterations are independent from each other allows to easily scale up the run time through parallelization on multiple machines, hence, in theory, allowing to speed up the calculation time to any desired time frame. And furthermore, despite this simulation being a 100 % sample, a common procedure is to only process fractions opposed to processing the entire population. By drawing a random sample, the results can then be extrapolated to reach the original level.

Axhausen claims that for models to be used by professionals during their every day work, run-times of no more than one night on standard computers are required [2]. While with the help of sub-samples and parallelization the runtime can be significantly decreased, the algorithm design itself also provides further potential. The most significant potential for speed improvements is that while the present search algorithm works

well for logbooks with up to five trips, the search is not optimized for logbooks with many trips. Logbooks with high trip numbers generally lead to large search spaces of potential combinations and as such, require a lot of computation time. Although only few logbooks exist with high trip numbers, they still make up a substantial part of the total run time. The current implementation shortens run time in that it terminates once a solution has been found without testing for other, perhaps more realistic solutions. Implementing a heuristic approach promises to shorten run times while obtaining more accurate results. The current implementation neither takes into account whether a query has been executed before. Hence, a caching strategy that avoids the multiple execution of identical search queries may lead to further optimization.

7 Discussion and conclusions

The previous chapter presented the development and data gathering process for the demonstration example and concluded with a discussion of the particular simulation results that were obtained. Continuing in this vein, a review of the method's key aspects from a broader context is the intention of this chapter. The first part summarizes the essential considerations that led to the methodology's design in section 7.1. The methodology's strengths and weaknesses are then assessed in section 7.2 and finally, the method's ability for forecasting is the focus of section 7.3.

7.1 Methodology design aspects

Just as with the previous example demonstration, most transport models simulate the period of a single day, usually one average work day. There also exist transport demand models that consider a complete week (examples are [4] and [30]), acknowledging the idea that the activities carried out in one day are affected by the activities carried out on other days [44]. The period of time that the present methodology emulates is determined by the time range that is covered by the template logbooks and can be one or several days. Time is represented in minutes, allowing the reproduction of features such as morning and evening peaks during rush hour.

The principle procedure of consulting a template logbook's geometric pattern for the distribution of trips as opposed to matching production and attraction key figures known from the traditional four-step process is shaped by the fact that meeting this model's spatial resolution requires production and attraction rates at the same level of detail, namely individual houses. Having to derive production and attraction rates from empirical data is avoided by instead referring to a template logbook's geometric shape.

Some transport demand models define capacity constraints (CC). The capacity constraint of a movie theater, for example, is the maximum number of customers that can be accommodated by it at one time. While capacity constraints at the level of zones and districts are essential for aggregated modeling approaches, the concept is not

adopted by most simulation-based implementations on the micro level [2]. One reason is that obtaining meaningful capacity constraints is costly and time-consuming. For instance, the capacities of movie theaters, schools, and restaurants need to be known or reasonable assumptions must be made if not. Another is the accompanied significant increase in computing time, if capacity constraints are to be considered. While the algorithm can be extended to evaluate capacity constraints without much programing effort, this feature has not been implemented for these reasons. Furthermore, the benefits that can be expected seem small in comparison to the accompanying costs; because of the way that trips are distributed in space, large queuing in certain locations may be less likely.

Spatial constraints were defined as "hard" constraints, meaning that they are applied in a strict sense. In contrast, the opposite would be letting the results violate some of the constraints or to differ to a certain degree. In the demonstration example that was given in chapter 6, waypoints of type one (terminal, station, port, airport) were defined to be located only in land-use areas of 'special use'. This can be problematic if, in the real world, not 100 % of such waypoints, but only a share is located in those areas. For example, only 80 % or 90 % of the waypoints of type one might actually be located in areas of special use. These and similar situations can be better expressed through random utility models (RUM). One advantage of RUMs is that they work with probabilities, opposed to a strict interpretation of hard constraints. As a consequence, they are more fail-safe with regard to inaccurate input data. In the light of the present methodology, the application of flexible template logbooks would further reduce the number of logbooks that cannot be matched, given that the more important constraints fit. Yet there are also drawbacks to be considered. A major disadvantage of random utility models is that they were originally intended for the description of small choice sets [54]. The application to large sets of alternatives would result in very large search spaces. Working with probabilities requires propagating the entire search tree, which is not feasable in reasonable computation time. On the contrary, the use of 'hard' constraints allows many branches of the search tree to be cut off, which is why the latter concept (i.e. 'hard' constraints) was implemented in the present demonstration example.

With respect to the two alternatives for modeling the waypoint selection procedure, the downsides of using 'hard' constraints can be reduced when generating the virtual world. The first alternative is to design the waypoint type constraint c_1 as outlayed in section 5.4.1, that is waypoints of the type ´shipping agency' are defined to be locations

in areas of industrial land-use housing a firm within the economic sector 'transport, storage, and communication':

$$\begin{aligned} C_{\text{shipping agency}} = \{c \in C \mid c \text{ 'is business'} \\ \wedge\, a_{\text{sector}}(c) = \text{'transport, storage, and communication'} \\ \wedge\, a_{\text{land-use}}(c) = \text{'industrial'}\} \end{aligned} \tag{7.1}$$

This waypoint type definition was chosen to differentiate between the waypoints of the various types by the characteristics that are provided by the spatial input data. On the contrary and as a second alternative, the location choice requirements may be considered at an early stage during the generation of the synthetic world. Given that the virtual world should resemble the real world as close as possible, waypoints of type 'shipping agency' should at best be defined as such and hence flagged accordingly. The constraint can then be simplified by changing the above definition (equation 7.1) into

$$C_{\text{shipping agency}} = \{c \in C \mid c \text{ 'is shipping agency'}\} \tag{7.2}$$

A benefit is that the number of shipping agency locations (to continue the example from above) can be predefined at the simulation's start, as opposed to no restrictions in the alternative one. Furthermore, for selective regions, spatial data exists in which not only the location of a building, but also the function it provides, is specified. When implementing data of that kind, trips to waypoints of a given type can be precisely defined as a trip to a location of that type. In the example demonstration in chapter 6, the first alternative was chosen because the data at hand did not allow for the second. Future implementations however, may benefit from a more detailed virtual world, especially in the light of upcoming improvements in data availability.

The methodology may also be applied to regions other than Berlin if the data requirements that were outlined in chapter 4.2 can be met. That is, data for the three categories trip chain data (section 4.2.1), traffic volume data (section 4.2.2), and population data (section 4.2.3) must be available. Areas containing major barriers to transport such as rivers and/or mountain chains will benefit from using the routed distance opposed to the airline distance, with respect to the effects on travel patterns. By matter of fact, the share of successfully matched logbooks can be expected to be high assuming that both the input data as well as the constraint definitions resemble

the behavior patterns of the study area. And further, the parameter epsilon allows for calibration.

7.2 Strengths and weaknesses

One strength of the methodology is that the demand generated by it is both sound on the macro as well as on the micro level. Results, when aggregated, are based on individual trip chains, which themselves are sound in that they reproduce behavior patterns that were empirically observed. As such, the output is of use to agent-based traffic flow simulators, which otherwise need different ways to disaggregate given aggregated trip numbers into individual trip chains. As research has mainly gone into improving the actual traffic flow models, very little attention has been given to the generation of trip chains, despite this being needed for input by these models. For this need, the methodology presented here improves current practice and provides a more sophisticated approach by consulting spatial and empirical data. The collection of trip chain data in particular will benefit from low surveying costs, supported by the trend towards personal GPS devices in many people's daily lives.

Another strength is that despite the principle intention to model transport demand for commercial passenger transport, the methodology is not limited to this type. The application to private households with behavior patterns typical to household members results in the demand for private trips and by generating the demand for freight transport likewise, a region's total transport demand can be obtained. For private transport, the simulation loop structure would then iterate over the set of households opposed to firms. Depending on the population's level of detail, the number of individuals in each household can be expressed by multiple mobility ratios per household to resemble the various person types that together constitute the household, e. g. the number of children or adults in work.

The methodology is further extendable in that it is not restricted to a given set of specific attributes. Peterson and Vovsha, for example, request that travel demand models should incorporate a structural typology of vehicles in the same way that they incorporate travel purposes [73]. Given that template logbooks are supplied with a specification of vehicle type, that information can be passed on to show up later in the output statistics. Depending on the empirical data at hand, this can be used to distinguish between the two types 'diesel' or 'gasoline' or to contribute to a more complex set of attributes such as vehicle type, emissions, age, and others.

There are also weaknesses. One is the current backtracking procedure, which always generates the same output for a given set of input parameters. For logbooks that contain two or more identical closed trip chains, the current implementation will cause the generated trips to be identical as well, while in reality the tours might have differed. The same applies if one template logbook is applied several times by a given firm. By selecting the first waypoint by random, this behavior can be circumvented in future implementations. Another improvement to the search algorithm would be to search for the match that best reflects the original template logbook's shape in that it minimizes the differences between original and matched distance-to-home values. In contrast, the current implementation terminates once the first valid solution has been found and does not search for subsequent solutions. Since doing so would require the iteration of the complete search tree, a heuristic approach should be chosen that finds optimal solutions in optimal time.

Further, the input data did not fit the data requirements perfectly. Recent and future improvements in data quality will allow more precise definitions of the simulation parameters in future implementations. The need for more accurate distance-to-home values has been acknowledged and will be answered by the upcoming remake of KiD, for which distance-to-home values will be provided, calculated prior to the data anonymization process. For the demonstration example in chapter 6, these values were calculated after the data anonymization necessary for data privacy compliance – in effect suboptimal distance values were obtained. Another improvement would be to survey more details on the waypoints reached in logbook surveys, such as the economic sector of external companies that are visited. In this regard, the remake of the survey KiD will lead to a significant improvement of the current situation by specifying the land-use at each waypoint.

Finally, the methodology is subject to the aspects that were identified in chapter 4.1, and to these affecting the capacity for behavior analysis, especially in the light of policy options that seek a change in people's route choice behavior. How this drawback can be resolved is the focus of the next section.

7.3 Scenario analysis and forecasting

Any attempt to make valid predictions about the future must be handled with care. For instance, one can argue that by matter of fact it is not possible to make valid predictions about the future. The philosopher Immanuel Kant claimed that any attempt to do so

must fail, unless it is the forecaster themself who has the power to take the actions needed for the prediction to become true:

> "A true forecast of a coming event is only possible if the forecaster is fully able to effect that which they priorly announced [48]." [1]

Nils Bohr, a Nobel Prize laureate in Physics, once joked that "prediction is very difficult, especially if it's about the future". In this, he is not alone. The statement is also known in one of its many variants from Mark Twain, Bernhard Shaw and Albert Einstein. On the other hand, predictions are a necessary and important tool for society. Despite being difficult and theoretically impossible, having actions rooted in the best knowledge available is by any means more favorable than taking decisions without any foundation. Similarily, no one seriously proposes investing in large-scale transport infrastructure projects without extensively investigating into the need for the investment. The principle scientific methodology that is typically applied is the use of models that forecast observed trends. Here, the quote by Bohr is a warning of the importance of finding a model that correctly identifies patterns in past data that will continue to hold true in the future [67]. A theoretical basis are ceteris paribus assumptions, that is to investigate into the effects of changing one parameter at a time, "with all other things being equal". The systematic approach allows making solid predictions based on arguable assumptions, and by that minimizes the risk of incorrect forecasts.

In the particular field of transport research, a number of scenarios are typically developed and tested against each other. Starting from a base scenario, policy options and the anticipated changes are evaluated. The presented methodology, in principle, allows the alteration of any of the input parameters as well as of the actual input data. Demographic, structural and economic changes can be depicted through the adaption of the virtual world:

1. First, manipulating the land-use data such as adding or removing industrial and residential land-use areas will affect the spatial distribution of trips.

2. Demographic changes can be accounted for by altering the virtual world's population accordingly, such as changing the grouping of individuals into households.

1 own translation from the German: Eine wahrsagende Geschichtsschreibung des Bevorstehenden in der künftigen Zeit ist nur möglich, wenn der Wahrsager die Begebenheiten selber macht und veranstaltet, die er zum voraus verkündigt.

Given that mobility ratios are known for the relevant household types, the effects can be measured.

3. For the integration of expected economic changes, the actual number of firms or the shares of firms per economic sector can be altered, such as shifts from one economic sector to another or smaller groups of large firms trending towards smaller firm sizes.

Changes in trip volume can be depicted by one of the following options:

1. The mobility ratios may be modified for scenarios in which the traffic volume is expected to change, such as changes in oil price, increases or losses in gross domestic product (GDP) or for relevant changes in the national unemployment numbers. This presumes that the expected changes in traffic volume must be determined beforehand through separate models. The present methodology then can be used to generate trip chain data based on the new input data assuming that the template logbooks embodied in the logbook repository remain unchanged.

2. Another way to model changes in trip volume is to manipulate the logbook repository in order to depict changes in tour patterns. For this purpose, logbooks can be generated synthetically by means of changing the distance of trips or the number of trips of some or all economic sectors, households or individuals.

While no technical difficulties hinder the joint application of a mix of measures, the ceteris paribus principle requires the application of only one of the above measures at a time. Of advantage is that the methodology is particularly helpful for scenarios in which only limited choice is given to the individuals, that is use cases in which the choice is not by the traveler, but other stakeholders who do not consider their decision's immediate impact on transport. Courier, express and parcel services (CEP) such as FedEx, DHL or UPS, are often constrained by customer demands that request the delivery of a shipment until a certain hour. The business model requests that shipments not be delivered prior the time that is payed for, in order to keep the added value of express deliveries exclusive.

8 Summary and outlook

8.1 Summary

Several conclusions can be made. Firstly, the demonstration showed that the methodology is applicable. The program code is maintainable and runs in reasonable time, allowing the simulation of 100 percent samples of large urban areas. The application also showed that much of the needed input data is available and the methodology can be adapted to fit new data, once available. Upcoming surveys will further lead to an increasing pool of data for the approach. For validation, the simulation results were sound in that they corresponded to third-party estimates. Of particular advantage is the high level of disaggregation and the ability to supply traffic flow simulators with agent plans in this detail.

The demonstration also identified weaknesses of the present approach. The input data did not fit the data requirements perfectly. Template logbooks from outside the study area were used to increase case numbers and the distance-to-home values in the template logbooks were affected by the prior data anonymization process. Another weakness is that the current implementation is not much optimized for processing speed. However, these weaknesses are not found in the principal design and can be averted in future implementations.

8.2 Outlook

During method implementation and demonstration, various aspects were identified throughout this thesis on how the existing method can be improved. As such, these considerations guide the way towards the next steps deemed necessary to establish the method as a standardized and easy to use tool. Thus a number of small issues were identified that can be easily taken into account in updated versions: epsilon should be expressed relative to the trip distance rather than absolutely; or to fine-tune the procedure for trips outside the synthetic world (e. g. by ranking possible waypoints according to their distance to the home site rather than choosing waypoints without

priority if the trip distance circle and the distance-to-home circle do not intersect). A significant improvement would be to replace the current location assignment procedure that is currently based on a trip's airline distance by the actual travel time. The benefit would be more realistic trip chains that fit well to local conditions, enabling the simulation of more regions. Another major improvement would be to search for trip chains that best reflect the template's spatial pattern, as in the current form, the algorithm moves on once the first valid solution is found. When addressing this issue, the algorithm's search strategy and run time may be optimized as well, something that was only partially covered in this thesis. And for cases in which the concept of capacity constraints is needed, the method may be extended accordingly.

Further prospects are the application to use cases that go beyond the original intention. While the generation of urban commercial motorized vehicle trip chains was the focus of this thesis, the method is not restricted to a given transport type or region. Hence, further research may test the method with other transport types and modes. And finally, linking the method with other micro simulation models promises further exploitation.

A Appendix

Table A.1: German Classification of Economic Activities, Edition 2003

WZ 2003	description
A	Agriculture, hunting and forestry
B	Fishing
C	Mining and quarrying
D	Manufacturing
E	Electricity, gas and water supply
F	Construction
G	Wholesale and retail trade; repair of motor vehicles, motorcycles and personal and household goods
H	Hotels and restaurants
I	Transport, storage and communication
J	Financial intermediation
K	Real estate, renting and business activities
L	Public administration and defence; compulsory social security
M	Education
N	Health and social work
O	Other community, social and personal service activities
P	Private households with employed persons
Q	Extra-territorial organizations and bodies

Source: Statistisches Bundesamt, Wiesbaden 2003 [85]

Listing A.1: The PL/SQL script that was used for creating the logbook repository from the empirical survey *Kraftfahrzeugverkehr in Deutschland* (KiD) for the demonstration example in chapter 6.

```
CREATE OR REPLACE FUNCTION "pwvm_createLogbooksFromKid"() RETURNS integer AS $$
DECLARE
    lb RECORD;
    trip RECORD;
    geom geometry;
BEGIN

        DELETE FROM pwvm_logbook_trip;
        DELETE FROM pwvm_logbook;

        FOR lb IN
        SELECT kid_fahrzeug.h01, f03::integer, kid_fahrzeug.fahrzeugid, f02c, f02a,
            kid_fahrt.f02_geom,
        (CASE WHEN kid_fahrzeug.h01='P' THEN null ELSE "pwvm_Theadcountclass".id END)
            AS headcountclass,
        kid_fahrzeug.k01::integer
        FROM kid_fahrzeug
        JOIN kid_fahrt ON (kid_fahrt.fahrzeugid = kid_fahrzeug.fahrzeugid)
        LEFT JOIN "pwvm_Theadcountclass"
        ON (kid_fahrzeug.h05 >= "pwvm_Theadcountclass".min AND kid_fahrzeug.h05 <= "
            pwvm_Theadcountclass".max)
        WHERE kid_fahrt.fahrtid = 1
    LOOP

        -- Now "lb" has one record from the query above


  INSERT INTO pwvm_logbook
    (id,
    source_type,
    "kid_vehicleId",
    source_plz,
    business_wz,
    business_headcountclass,
    source_kgs,
    the_geom,
    vehicletype)
  VALUES (lb.fahrzeugid, lb.f03, lb.fahrzeugid, lb.f02c, lb.h01, lb.headcountclass, lb.
      f02a, lb.f02_geom, lb.k01);
```

```
-- logbookId = fahrzeugid

geom = lb.f02_geom;

FOR trip IN
      SELECT fahrtid, f09::integer, f08c, f08a, f07a::integer, f14::double precision,
           f08_geom, f04, f10a
      FROM kid_fahrt
      WHERE fahrzeugid = lb.fahrzeugid
      ORDER BY fahrtid ASC
      LOOP

      INSERT INTO pwvm_logbook_trip ("logbookId", "tripId", dest_type, dest_plz,
          dest_kgs, purpose, distance_empirical, the_geom, start_time, stop_time)
      VALUES (
      lb.fahrzeugid,
      trip.fahrtid,
      trip.f09,
      trip.f08c,
      trip.f08a,
      trip.f07a,
      trip.f14,
      trip.f08_geom,
      (CASE WHEN trip.f04='-1:-1' OR trip.f04='-9:-9' THEN null ELSE CAST(trip.f04 AS
           TIME WITHOUT TIME ZONE) END),
      (CASE WHEN trip.f10a='-1:-1' OR trip.f10a='-9:-9' THEN null ELSE CAST(trip.f10a
           AS TIME WITHOUT TIME ZONE) END)
      );

      IF lb.f02_geom IS NOT NULL AND trip.f08_geom IS NOT NULL
      THEN
        UPDATE pwvm_logbook_trip SET z_source = ST_distance(lb.f02_geom, trip.
            f08_geom)
        WHERE "logbookId" = lb.fahrzeugid AND "tripId" = trip.fahrtid;
      END IF;

      IF geom IS NOT NULL AND trip.f08_geom IS NOT NULL
      THEN
        UPDATE pwvm_logbook_trip SET distance = ST_distance(geom, trip.f08_geom)
        WHERE "logbookId" = lb.fahrzeugid AND "tripId" = trip.fahrtid;
      END IF;
```

```
      geom = trip.f08_geom;

END LOOP;

-- Resolving branch office vs. own business

-- If multiple waypoints are of type 'home business', all are set to be of type '
    branch office' except:
-- by priority:
-- 1. The waypoint that appears more frequently than others and that is identified
    as 'home business' in all cases
-- 2. The first waypoint of type 'home business'

UPDATE pwvm_logbook
SET source_type = 9 -- branch office
WHERE id = lb.fahrzeugid
AND source_type = 4 -- home business
AND the_geom != (
  SELECT the_geom FROM (
    SELECT count(*), type, the_geom FROM (
      SELECT the_geom, source_type AS type FROM pwvm_logbook WHERE id = lb.fahrzeugid
        UNION
      SELECT the_geom, dest_type FROM pwvm_logbook_trip WHERE "logbookId" = lb.
          fahrzeugid
      ) waypointlist
      WHERE type = 4 -- home business
      GROUP BY the_geom, type
      ORDER BY count DESC
    ) hb
  LIMIT 1
);

UPDATE pwvm_logbook_trip
SET dest_type = 9 -- Filiale
WHERE "logbookId" = lb.fahrzeugid
AND dest_type = 4 -- home business
AND the_geom != (
  SELECT the_geom FROM (
    SELECT count(*), type, the_geom FROM (
      SELECT the_geom, source_type AS type FROM pwvm_logbook WHERE id = lb.fahrzeugid
        UNION
      SELECT the_geom, dest_type FROM pwvm_logbook_trip WHERE "logbookId" = lb.
          fahrzeugid
```

```
      ) waypointlist
      WHERE type = 4 -- home business
      GROUP BY the_geom, type
      ORDER BY count DESC
    ) hb
  LIMIT 1
);

  -- END Resolving branch office vs. own business


END LOOP;

  -- Generating z

UPDATE pwvm_logbook_trip t
SET z_home =
  CASE
    WHEN t.the_geom IS NOT NULL AND (SELECT the_geom FROM pwvm_logbook_trip WHERE
        pwvm_logbook_trip."logbookId"=t."logbookId" AND pwvm_logbook_trip.dest_type =
         4 LIMIT 1) IS NOT NULL
    THEN ST_distance((SELECT the_geom FROM pwvm_logbook_trip current_trip WHERE
        current_trip."logbookId"=t."logbookId" AND current_trip.dest_type = 4 LIMIT
        1), t.the_geom)
    ELSE
      CASE
        WHEN t.the_geom IS NOT NULL AND (SELECT the_geom FROM pwvm_logbook WHERE id=t.
            "logbookId") IS NOT NULL AND (SELECT source_type FROM pwvm_logbook WHERE
            id=t."logbookId") = 4
        THEN ST_distance((SELECT the_geom FROM pwvm_logbook WHERE id=t."logbookId"),
            t.the_geom)
      END
 END;



-- Filter according to trip purpose
DELETE FROM pwvm_logbook_trip WHERE "pwvm_getNumberOfBusiness
    PassengerTransportTrips"("logbookId") <= 0;

-- Delete trips for which the covered airline distance could not be determined
DELETE FROM pwvm_logbook_trip WHERE "logbookId" IN (SELECT "logbookId" FROM
    pwvm_logbook_trip WHERE distance IS NULL);
```

```
  -- Delete records in which multiple trips are summarized into one trip
  DELETE FROM pwvm_logbook_trip WHERE "logbookId" IN (SELECT fahrzeugid FROM kid_fahrt
       WHERE f16 = '2');

  -- clean up (ensuring that only logbooks are kept with at least one trip)
  DELETE FROM pwvm_logbook WHERE id NOT IN (SELECT "logbookId" FROM pwvm_logbook_trip);


  RETURN 1;

END;
$$ LANGUAGE plpgsql;
```

Table A.2: Demonstration example: number of firms in the study area

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1-9 employees	780		24	4449	268	4312	21565	4063	4005	3508	27097	750	2739	11755	18641
10-49 employees	81		23	2059	79	2606	1032	194	222	109	1629	239	634	910	670
50-99 employees	5			260	6	226	114	24	18	9	86	9	51	41	46
100-249 employees	5			79	1	38	65	18	14	4	82	6	17	27	52
250-499 employees	2			46	2	8	26	8	6	12	37	9	3	26	35
500-1000 employees				10		4		1	2	1	7	2		3	3
$>$ 1000 employees				8	3		3		2	2	2	3	1	1	2
not specified	204		65	2826	152	1243	3849	869	921	2044	7214	105	595	1360	5435
sum	1077		112	9737	511	8437	26654	5177	5190	5689	36154	1123	4040	14123	24884

Source: infas Firmenzaehler, see also [60]. See table A.1 for a description of all economic sectors.

Table A.3: Demonstration example: number of visits to customers by firms, by economic sector and number of employees

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1-9 employees				0.33	0.16	1	0.5		4.93	3.59	1.41		2.98		1.19
10-49 employees		0.01		0.66	0.01	0	6.34		8.27	3.38	8.63		0.4	0.95	0.17
50-99 employees			0.09	0.82	0.12	19.63	88.61		1.4	7.43	25.65		1.38		48.92
100-249 empl.				63.1	1.31	15.29	1.38		280.1	85.11	24.15			7.14	16.17
250-499 empl.			0.14	0.28	23.48	222.52	31.09	10.09	0.02	9.54	35.61	190.14			7.09
500-1000 empl.			4.38	0.05			0.04		0	6.23	126.11		0.08	0.02	5.95
$>$ 1000 employees			0.02				0		0.16	453.57	13.13	0.02			204.29

Source: Based on information from the empirical survey *Dienstleistungsverkehr in industriellen Wertschöpfungsprozessen* [43]. See table A.1 for a description of all economic sectors.

Table A.4: Demonstration example: average number of visits to customers per KiD logbook, by economic sector and number of employees. Not all numbers are statistically sound due to the low number of cases.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1-9 employees	1.52	0.67	2.08	2.03	1.93	1.77	2.22	1.07	2.84	1.86	1.8	2.23	1.06	2.59	1.98
10-49 employees	1.43	4	1.04	1.67	2.35	1.69	1.99	1.26	2.32	0.9	1.83	2.04	1.82	3.12	2.23
50-99 employees	1	0	2	1.56	2.42	1.61	2.07	1.43	2.12	1.31	2.24	2.24	0.83	3.51	2
100-249 employees	1.5	0	0.85	1.66	2.59	1.83	2.23	1.75	2.13	1.13	1.69	1.49	2.17	2.1	1.6
250-499 employees	1	0	1	1.95	2.67	1.49	2.1	2.5	1.43	0.86	2.24	2.16	2	2.76	1.6
500-1000 employees	0	0	1	1.61	1.87	1.83	1.43	2	2.36	1.1	1.82	2.08	1.1	1.45	1.8
$>$ 1000 employees	0	0	0.4	1.85	2.2	1.5	1.2	3.25	1.43	1.56	1.14	1.86	0	1.8	2.5
n.s.	1.59	0	1.14	2.16	2.31	1.67	2.47	3.25	2.16	1.14	2.17	1.6	1.1	2.3	2.18

Source: *Kraftfahrzeugverkehr in Deutschland* (KiD) [100]. See table A.1 for a description of all economic sectors.

Bibliography

[1] Amt für Statistik Berlin–Brandenburg. Fortgeschriebene Bevölkerungszahl vom 31. Mai 2010. Website. www.statistik-berlin-brandenburg.de, 2010. Accessed on Oct. 26th.

[2] Kay W. Axhausen. Neue Modellansätze der Verkehrsnachfragesimulation. Entwicklungslinien, Stand der Forschung, Forschungsperspektiven. Technical report, Institut für Verkehrsplanung und Transportsysteme (IVT), Eidgenössische Technische Hochschule Zürich (ETH), 2006.

[3] Kay W. Axhausen and Lorenz Hurni, editors. *Zeitkarten der Schweiz*. Institut für Verkehrsplanung und Transportsysteme (IVT) and Institut für Kartographie (IKA), Eidgenössische Technische Hochschule Zürich (ETH), 2005.

[4] Kay W. Axhausen, Paul Widmer, and Ruedi Ott. Aktivitätenorientierte Personenverkehrsmodelle (Vorstudie). In *Arbeitsbericht Verkehrs- und Raumplanung*, volume 70. Institut für Verkehrsplanung und Transportsysteme (IVT), Eidgenössische Technische Hochschule Zürich (ETH), 2001.

[5] Martin Bach, Lutz Breuer, Hans-Georg Frede, J.A. Huisman, Annette Otte, and Rainer Waldhardt. Accuracy and congruency of three different digital land-use maps. *Landscape and Urban Planning*, 78(4):289 – 299, 2006.

[6] Michael Balmer. Travel demand generation for multi-agent transport simulations: algorithms and systems. Doctoral thesis, Eidgenössische Technische Hochschule Zürich (ETH), Switzerland, 2007.

[7] Michael Balmer, Kay W. Axhausen, and Kai Nagel. An agent-based demand-modeling framework for large scale micro-simulations. *Transportation Research Board, 85th Annual Meeting*, 2006.

[8] Richard J. Beckman, Keith A. Baggerly, and Michael D. McKay. Creating synthetic baseline populations. *Transportation Research Part A: Policy and Practice*, 30(6):415–429, 1996.

[9] Moshe Ben-Akiva, Jon Bottom, Song Gao, Haris N. Koutsopoulos, and Yang Wen. Towards disaggregate dynamic travel forecasting models. *Tsinghua Science and Technology*, 12(2):115–130, 2007.

[10] Itzhak Benenson and Paul M. Torrens. *Geosimulation – Automata-based modeling of urban phenomena*. John Wiley & Sons, 2004.

[11] Ulrike Beuck, Kai Nagel, and Andreas Justen. Application of the VISEVA demand generation software to Berlin using publicly available behavioral data. *Transportation Research Board, 85th Annual Meeting*, 2006.

[12] Ulrike Beuck, Marcel Rieser, David Strippgen, Michael Balmer, and Kai Nagel. Preliminary results of a multi-agent traffic simulation for Berlin. In Sergio Albeverio, Denise Andrey, Paolo Giordano, and Alberto Vancheri, editors, *The Dynamics of Complex Urban Systems*, pages 75–94. Physica–Verlag HD, 2008. 10.1007/978-3-7908-1937-3_5.

[13] Anette Blaser. ALKIS, das Amtliche Liegenschaftskataster Informationssystem: Von Karte und Buch zum integrierten System. In *Geoinformation 2010 in Berlin: Veränderungen und Herausforderungen*, pages 57–59. Senatsverwaltung für Stadtentwicklung, November 2010.

[14] Rupert Bobinger. Modellierung der verkehrsnachfrage bei preispolitischen massnahmen. Doctoral thesis, Techniche Universtität München, Germany, 2000.

[15] John L. Bowman. The day activity schedule approach to travel demand analysis. Doctoral thesis, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA, 1998.

[16] John L. Bowman. Historical development of activity based model theory and practice. *Traffic Engineering and Control*, Vol. 50 No. 2: 59–62 (part 1), Vol. 50 No. 7: 314–318 (part 2), 2009.

[17] John L. Bowman. Population synthesizers. *Traffic Engineering and Control*, 49(9):342, 2009.

[18] John L. Bowman and Moshe E. Ben-Akiva. Activity-based disaggregate travel demand model system with activity schedules. *Transportation Research Part A: Policy and Practice*, 35(1):1 – 28, 2001.

[19] Werner Brög and Gerhard Winter. Untersuchungen zum Problem der "non-reported-trips" zum Personen-Wirtschaftsverkehr bei Hautshaltsbefragungen. *Schriftenreihe Forschung Straßenbau und Straßenverkehrstechnik*, Heft 593, 1990.

[20] Bundesamt für Kartographie und Geodäsie (BKG). Georeferenzierte Adressdaten – Bund (GAB), 2008.

[21] Bundesamt für Kartographie und Geodäsie (BKG). Website. www.bkg.bund.de, 2010. Accessed on May 25th.

[22] Howard Butler, Christopher Schmidt, Dane Springmeyer, and Josh Livni. Spatial Reference. Website. www.spatialreference.org, 2009. Accessed on July 23rd.

[23] Charles W. Cobb and Paul H. Douglas. A theory of production. *American Economic Review*, 18(1 (Supplement)):139–165, 1928.

[24] D. Damm. Theory and empirical results: a comparison of recent activity-based research. In S. Carpenter and P.M. Jones, editors, *Recent Advances in Travel Demand Analysis*, pages 3–33. Gower, Aldershot, England, 1983.

[25] Data Documentation Initiative (DDI) Alliance. Welcome to the Data Documentation Initiative. Website. www.ddialliance.org, 2010. Accessed on Oct. 13th.

[26] Juan de Dios Ortuzar and Luis G. Willumsen. *Modelling Transport*. Wiley & Sons, 3rd edition, 2001.

[27] William E. Deming and Frederick F. Stephan. On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics*, Vol. 11(No. 4):427–444, 1940.

[28] Kai M. Deneke. Nutzungsorientierte Fahrzeugkategorien im Straßenwirtschaftsverkehr – Eine multidimensionale Analyse kraftfahrzeugbezogener Mobilitätsstrukturen. Doctoral thesis, Technische Universtität Braunschweig, Germany, 2004.

[29] Deutsches Zentrum für Luft- und Raumfahrt (DLR). SUMO: Simulation of Urban MObility. Website. http://sumo.sourceforge.net, 2011. Accessed on Feb. 7th.

[30] Sean T. Doherty and Kay W. Axhausen. The development of a unified modeling framework for the household activity-travel scheduling process. In W. Brilon,

F. Huber, M. Schreckengerg, and H. Wallentowitzpp, editors, *Traffic and Mobility: Simulation – Economics – Environment*, pages 35–36. Springer, 1999.

[31] Eurostat. NACE Rev. 2 Statistical classification of economic activites in the European Community. *Methodologies and Working papers*, 2008.

[32] Eurostat. Nomenclature of Statistical Territorial Units (NUTS). Website. http://epp.eurostat.ec.europa.eu/portal/page/portal/nuts_nomenclature/introduction, 2011. Accessed on Feb. 7th.

[33] Bent Flyvbjerg. Measuring inaccuracy in travel demand forecasting: methodological considerations regarding ramp up and sampling. *Transportation Research Part A: Policy and Practice*, 39(6):522 – 530, 2005.

[34] Bent Flyvbjerg, Mette K. Skamris, and Søren L. Buhl. Inaccuracy in traffic forecasts. *Transport Reviews*, 26(1):1–24, 2006.

[35] Robert Follmer, Dana Gruschwitz, Birgit Jesske, Sylvia Quandt, Barbara Lenz, Claudia Nobis, Katja Köhler, and Markus Mehlin. Mobilität in Deutschland 2008. Ergebnisbericht: Struktur – Aufkommen – Emissionen – Trends, 2010. Online at www.mobilitaet-in-deutschland.de/pdf/MiD2008_Abschlussbericht_I.pdf.

[36] Roman Frigg and Stephan Hartmann. *Models in Science*. The Stanford Encyclopedia of Philosophy (Spring 2006 Edition), 2008. Online at http://plato.stanford.edu/entries/models-science/.

[37] Nigel Gilbert. *Agent-based Models*, volume 153 of *Quantitative Applications in the Social Sciences*. SAGE Publications, 2008.

[38] Marta C. Gonzalez, Cesar A. Hidalgo, and Albert-Laszlo Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, June 2008.

[39] Google Inc. Google Maps. Website. http://maps.google.de, 2010. Accessed on Oct. 19th.

[40] Reinhard Gressel and Christophe Mundutéguy. Les professionnels mobiles: Un groupe hétérogène avec une exposition importante au risque routier. *RTS : Recherche Transports Sécurité*, 99:147–167, 2008.

[41] Torsten Hägerstrand. What about people in regional science? *Papers in Regional Science*, 24:6–21, 1970.

[42] Stephan Hartmann. The world as a process. Simulations in the natural and social sciences. In R. Hegselmann, U. Müller, and K. G. Troitzsch, editors, *Modelling and Simulation in the Social Sciences from the Philosophy of Science Point of View*, pages 77–100. Kluwer, 1996.

[43] Heinz Hautzinger, Marcus Bäumer, Anna Hussinger, Barbara Lenz, Julius Menge, Kurt Monse, Michael Fromm, Marco Andres, Tanja Klostermann, Frank Straube, and Wolf-Christian Hildebrand. Dienstleistungsverkehr in industriellen Wertschöpfungsprozessen – Schlussbericht. Final project report, 2008.

[44] Georg Hertkorn. Mikroskopische Modellierung von zeitabhängiger Verkehrsnachfrage und von Verkehrsflußmustern. Doctoral thesis, Universität zu Köln, Germany, 2004.

[45] Jinxing Hu, Wenjing Cao, Jun Luo, and Xiaomin Yu. Dynamic modeling of urban population travel behavior based on data fusion of mobile phone positioning data and FCD. In *17th International Conference on Geoinformatics, Aug. 12-14, Fairfax, USA*, pages 1–5, 2009.

[46] Eddie Hunsinger. Iterative proportional fitting for a two-dimensional table. Website. www.demog.berkeley.edu/~eddieh/IPFDescription/AKDOLWDIPFTWOD.pdf, 2008. Accessed on Dec. 21st, 2010.

[47] Peter M. Jones. *New approaches to understanding travel behaviour: the human activity approach*, chapter 2, pages 55–80. Croom Helm, 1979.

[48] Immanuel Kant. Erneuerte Frage: Ob das menschliche Geschlecht im beständigen Fortschreiten zum Besseren sei? (1789) cited after: Immanuel Kant: Von den Träumen der Vernunft. In *Kleine Schriften zur Kunst, Philosophie, Geschichte und Politik,*. Kiepenheuer, Leipzig/Weimar, 1979.

[49] Ryuichi Kitamura. An evaluation of activity-based travel analysis. *Transportation*, 15(1-2):9–34, 1988.

[50] Sigrid Klein-Vielhauer. Neue Konzepte für den Wirtschaftsverkehr in Ballungsräumen. Ein Werkstattbericht über Bemühungen in Praxis und Wissenschaft. *Wissenschaftliche Berichte/Forschungszentrum Karlsruhe, Technik und Umwelt (FZKA)*, 6599, 2001.

[51] Kraftfahrt-Bundesamt (KBA). Zentrales Fahrzeugregister. Website. www.kba.de/nn_261078/EN/ZentraleRegister__en/ZFZR__en/zfzr__inhalt__en.html, 2010. Accessed on May 25th.

[52] Stefan Krauß. Microscopic modeling of traffic flow: Investigation of collision free vehicle dynamics. Doctoral thesis, Universität zu Köln, Germany, 1998.

[53] Jens Landmann. Aufbereitung und Untersuchungen der Erhebungsdaten "Kraftfahrzeugverkehr in Deutschland" zur Nutzung für verkehrsplanerische Berechnungen. Diploma thesis, Technische Universität Dresden, Germany, 2005.

[54] Fabrice Marchal and Kai Nagel. Modeling location choice of secondary activities with a social network of cooperative agents. *Transportation Research Record*, 1935:141–146, 2005.

[55] MATSIM-T. Multi-Agent Transport Simulation Toolkit. Website. www.matsim.org, 2010. Accessed on Feb. 7th.

[56] Michael G. McNally. *Transport Modelling*, chapter The Four Step Model. Pergamon, Elsevier Science, 2nd edition, 2007.

[57] Michael G. McNally and Craig Rindt. *Transport Modeling*, chapter The Activity-Based Approach. Pergamon, Elsevier Science, 2nd edition, 2007.

[58] Markus Mehlin and Wiebke Zimmer. Ein Weg für klimagerechte Mobilität. *Internationales Verkehrswesen*, 62(1):10–15, 2010.

[59] Julius Menge and Marcus Bäumer. Personenwirtschaftsverkehr – Mobil im Namen der Dienstleistung? In Uwe Clausen, editor, *Wirtschaftsverkehr 2007*, page 115. Verlag Praxiswissen, 2007.

[60] Julius Menge, Clemens Bochynek, Sebastian Schneider, and Madlen Venus. Erstellung und Verwendung einer synthetischen Wirtschaftsstruktur zur disaggregierten Modellierung der Wirtschaftsverkehrsnachfrage. *Logistik, Verkehr und Umwelt*, pages 23–37, Juni 2009.

[61] Julius Menge and Barbara Lenz. Services and service traffic: A reappraisal from the perspective of transportation geography. In Ullrich Martin, editor, *Networks for Mobility 2008, 4th International Symposium*, 2008.

[62] Michael Meyer and Eric J. Miller. *Urban Transportation Planning*. McGraw-Hill, 2nd edition, 2000.

[63] Microsoft Corporation. Microsoft SQL Server 2008: Spatial Data. Website. www.microsoft.com/sqlserver/2008/en/us/spatial-data.aspx, 2009. Accessed on Jan. 21st.

[64] Michael Morgenstern. Unbalanced Germany. *The Economist*, August 6th, 2009.

[65] MySQL AB. MySQL 6.0 Reference Manual – Spatial Extensions. Website. http://dev.mysql.com/doc/refman/6.0/en/spatial-extensions.html, 2009. Accessed on Jan. 21st.

[66] Kai Nagel and Fabrice Marchal. Computational methods for multi-agent simulations of travel behavior. In *Moving through nets: The physical and social dimensions of travel; 10th International Conference on Travel Behaviour Research; Lucerne, Switzerland*, 2003.

[67] Bob Nau. Famous forecasting quotes. www.duke.edu/rnau/411quote.htm, 2011. Accessed on July 18th.

[68] Öko-Institut e.V. and Deutsches Zentrum für Luft- und Raumfahrt (DLR). RENEWBILITY – Stoffstromanalyse nachhaltige Mobilität im Kontext erneuerbarer Energien bis 2030 – Zentrale Ergebnisse. Report on project FZK 0327546, 2009.

[69] Open GIS Consortium, Inc. Welcome to the OGC. Website. www.opengeospatial.org, 2009. Accessed on Jan. 20th.

[70] Open GIS Consortium, Inc. (OGC). OpenGIS Simple Features Specification For SQL. revision 1.1, opengis project document 99-049, 1999. Online at http://portal.opengeospatial.org/files/?artifact_id=829.

[71] Oracle Corporation. Oracle spatial, locator, and location-based services. Website. www.oracle.com/technetwork/database/options/spatial/, 2011. Accessed on Feb. 7th.

[72] Paul M. Torrens. Geosimulation: innovative geospatial research. Website. www.geosimulation.org/geosim/, 2009. Accessed on Jan. 20th.

[73] Eric Petersen and Peter Vovsha. Intra-household car type choice for different travel needs. *Transportation Research Record*, pages 207–219, 2006.

[74] Alban W. Phillips. Mechanical models in economic dynamics. In *Economica 17*, pages 283–305, 1950.

[75] Carl Pirath. *Die Grundlagen der Verkehrswirtschaft*, volume 2. erweiterte Auflage. Springer, 1949.

[76] PTV AG. VALIDATE – Nationales Verkehrsmodell Deutschland, 2006.

[77] Bryan Keith Raney. Learning framework for large-scale multi-agent simulations. Doctoral thesis, Eidgenössische Technische Hochschule Zürich (ETH), Switzerland, 2005.

[78] Refractions Research. PostGIS. Website. www.postgis.org, 2011. Accessed on Feb. 7th.

[79] Thomas F. Rossi, Brad Winkler, Tim Ryan, Karen Faussett, Yali Li, Donna Wittl, and Maya Abou Zeid. Deciding on Moving to Activity-Based Models (or Not). *Transportation Research Board, 88th Annual Meeting*, 2009.

[80] Christian Schiller. Integration des ruhenden Verkehrs in die Verkehrsangebots- und Verkehrsnachfragemodellierung. Doctoral thesis, Technische Universität Dresden, Germany, 2004.

[81] Franz P. Schütte. Mobilitätsprofile im städtischen Personenwirtschaftsverkehr. Doctoral thesis, Technische Universität Dortmund, Germany, 1995.

[82] Senatsverwaltung für Stadtentwicklung. mobil2010. Stadtentwicklungsplan Verkehr Berlin, 2003. Online at www.stadtentwicklung.berlin.de/planen/stadtentwicklungsplanung/de/verkehr/download/Mobil2010.zip.

[83] Herbert Stachowiak. *Allgemeine Modelltheorie*. Springer, Wien[u.a.], 1973.

[84] Statistikregistergesetz (StatRegG). Gesetz zur Durchführung der Verordnung (EWG) Nr. 2186/93 des Rates vom 22. Juli 1993 über die innergemeinschaftliche Koordinierung des Aufbaus von Unternehmensregistern für statistische Verwendungszwecke. BGBl. I S. 1300, 2903, 1998.

[85] Statistisches Bundesamt. German classification of economic activity, 2003.

[86] Stephen V. Stehman and Raymond L. Czaplewski. Design and analysis for thematic map accuracy assessment: Fundamental principles. *Remote Sensing of Environment*, 64(3):331 – 344, 1998.

[87] Imke Steinmeyer. Kenndaten der Verkehrsentstehung im Personenwirtschaftsverkehr. Doctoral thesis, Technische Universität Hamburg-Harburg, Germany, 2004.

[88] Imke Steinmeyer. Definition und Bedeutung des Personenwirtschaftsverkehrs — Ein Sachstandsbericht aus dem Jahr 2006. *Schriften des Fachgebiets für Integrierte Verkehrsplanung an der TU Berlin*, 2007.

[89] Imke Steinmeyer and Tina Wagner. Verwendung der "Kraftfahrzeugverkehr in Deutschland" (KiD 2002) für städtische bzw. regionale Fragestellungen. *Wirtschaftsverkehr 2003. Trends – Modelle – Konzepte*, 2003.

[90] Imke Steinmeyer and Tina Wagner. Using national behavioral data on commercial traffic for local and regional applications: Experiences from Germany with data sources and gaps, opportunities and limits. *Transportation Research Board, 85th Annual Meeting*, 2006.

[91] David Strippgen. Investigating the technical possibilities of real-time interaction with simulations of mobile intelligent particles. Doctoral thesis, Technische Universität Berlin, Germany, 2009.

[92] David Strippgen and Kai Nagel. Using common graphics hardware for multi-agent traffic simulation with cuda. In *Simutools '09: Proceedings of the 2nd International Conference on Simulation Tools and Techniques*, pages 1–8, Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering (ICST), Brussels, Belgium, 2009.

[93] The RailML Initiative. railML.org. Website. www.railml.org, 2010. Accessed on Oct. 13th.

[94] TRANSIMS. TRansportation ANalysis and SIMulation System. Website. http://transims-opensource.net, 2010. Accessed on Feb. 7th.

[95] United Nations Statistics Division. International Standard Industrial Classification of All Economic Activities ISIC Rev. 4 – Detailed structure and explana-

tory notes, 2010. Online at http://unstats.un.org/unsd/cr/registry/regcst.asp?Cl=27&Lg=1.

[96] Verband Deutscher Städtestatistiker auf der Grundlage der geltenden Empfehlungen des Deutschen Städtetags. Kommunale Gebietsgliederung. Empfehlungen zur Ordnung des Strassen-/Hausnummernsystems und Gliederung des Gemeindegebiets nach Gemeindeteilen, Blöcken und Blockseiten sowie DV-Organisation. *DST-Beiträge zur Informationsgesellschaft und Stadtforschung*, Heft 39, 1991.

[97] Peter Wagner and Kai Nagel. Microscopic modelling of travel demand: The home-to-work problem. *Transportation Research Board, 78th Annual Meeting*, 1999.

[98] Volker Walter. Object-based classification of remote sensing data for change detection. *ISPRS Journal of Photogrammetry and Remote Sensing*, 58(3-4):225 – 238, 2004.

[99] John G. Wardrop. Some theoretical aspects of road traffic research. *Proceedings of the Institute of Civil Engineers*, 1:325–378, 1952.

[100] Manfred Wermuth. Kontinuierliche Befragung des Wirtschaftsverkehrs in unterschiedlichen Siedlungsräumen – Phase 2, Hauptstudie (Kraftfahrzeugverkehr in Deutschland – KiD 2002). Final report for project 70.0682/2001, 2003.

[101] Manfred Wermuth and Horst H. Binnenbruck. *Bestandsaufnahme notwendiger und verfügbarer Daten zum Wirtschaftsverkehr als Grundlage pragmatischer Datenergänzungen. Bericht zum Forschungs- und Entwicklungsvorhaben 01.145G96C des Bundesministeriums für Verkehr, Bau und Wohnungswesen.* Typo Druck, 2003.

[102] Manfred Wermuth, Christian Neef, and Imke Steinmeyer. Goods and business traffic in germany. In Peter Stopher and Cheryl C. Stecher, editors, *Travel Survey Methods – Quality and Future Directions*, pages 427–450. Elsevier, 2006.

[103] Rico Wittwer. Raumstrukturelle Einflüsse auf das Verkehrsverhalten. Doctoral thesis, Technische Universität Dresden, Germany, 2008.

[104] Working Committee of the Surveying Authorities of the States of the Federal Republic of Germany (Arbeitsgemeinschaft der Vermessungsverwaltungen

der Länder der Bundesrepublik Deutschland, AdV). Amtliches topographisch-kartographisches Informationssystem (ATKIS) – Gesamtdokumentation, 2002.

[105] Working Committee of the Surveying Authorities of the States of the Federal Republic of Germany (Arbeitsgemeinschaft der Vermessungsverwaltungen der Länder der Bundesrepublik Deutschland, AdV). ATKIS – Objektartenkatalog Basis-DLM, 2003. Version 3.2.

Acknowledgments

This work has been written at the German Aerospace Center (DLR) at the Institute of Transport Research (VF). I would like to thank my advisors Prof. Dr. Barbara Lenz, director of DLR-VF, and Prof. Dr. Kai Nagel from Technische Universität Berlin for their guide and help.

I would also like to thank Prof. Dr. Christoph Meinel, Scientific Director and CEO of Hasso-Plattner-Institut for IT-Systems Engineering (HPI), Prof. Dr. Bernd Walter, chair of the Database and Information Systems unit at Universität Trier, and Dr. Charles Petry from Stanford University. All of them substantially guided my way into academics.

Further I want to thank all colleagues of DLR-VF for their support and the great atmosphere at work, particularly Andreas Lischke and PD Dr.-Ing. habil. Christian Schiller for reviewing this work.

Enrico Zattarin and John Cole were exceptionally helpful with all aspects to language.