# D4.1 Development of a Novel Evaluation and Benchmarking Standard

## DATA science for SIMulating the era of electric vehicles

www.datasim-fp7.eu

<table>
<thead>
<tr>
<th>Project details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Project reference: 270833</td>
</tr>
<tr>
<td>Status: Execution</td>
</tr>
<tr>
<td>Programme acronym: FP7-ICT (FET Open)</td>
</tr>
<tr>
<td>Subprogramme area: ICT-2009.8.0 Future and Emerging Technologies</td>
</tr>
<tr>
<td>Contract type: Collaborative project (generic)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Consortium details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coordinator: 1. (UHasselt) Universiteit Hasselt</td>
</tr>
<tr>
<td>Partners: 2. (CNR) Consiglio Nazionale delle Ricerche</td>
</tr>
<tr>
<td>3. (BME) Budapesti Muszaki es Gazdasag tudomanyi Egyetem</td>
</tr>
<tr>
<td>4. (Fraunhofer) Fraunhofer-Gesellschaft zur Foerderung der Angewandten Forschung E.V</td>
</tr>
<tr>
<td>5. (UPM) Universidad Politecnica de Madrid</td>
</tr>
<tr>
<td>6. (VITO) Vlaamse Instelling voor Technologisch Onderzoek N.V.</td>
</tr>
<tr>
<td>7. (IIT) Technion – Israel Institute of Technology</td>
</tr>
<tr>
<td>8. (UPRC) University of Piraeus Research Center</td>
</tr>
<tr>
<td>9. (HU) University of Haifa</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Contact details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prof. dr. Davy Janssens</td>
</tr>
<tr>
<td>Universiteit Hasselt– Transportation Research Institute (IMOB)</td>
</tr>
<tr>
<td>Function in DATA SIM: Person in charge of scientific and technical/technological aspects</td>
</tr>
<tr>
<td>Address: Wetenschapspark 5 bus 6</td>
</tr>
<tr>
<td>Tel.: +32 (0)11 26 91 28</td>
</tr>
<tr>
<td>Fax: +32 (0)11 26 91 99</td>
</tr>
<tr>
<td>E-mail: <a href="mailto:davy.janssens@uhasselt.be">davy.janssens@uhasselt.be</a></td>
</tr>
<tr>
<td>URL: <a href="http://www.imob.uhasselt.be">www.imob.uhasselt.be</a></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Deliverable details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Work Package: 4</td>
</tr>
<tr>
<td>Deliverable: D4.1</td>
</tr>
<tr>
<td>Dissemination level: xxx</td>
</tr>
<tr>
<td>Nature: xxx</td>
</tr>
<tr>
<td>Contractual Date of Delivery: 31.08.2013</td>
</tr>
<tr>
<td>Actual Date of Delivery: 31.08.2013</td>
</tr>
<tr>
<td>Total number of pages: 42</td>
</tr>
<tr>
<td>Authors: Daniel Schulz, Bruno Kochan, Christine Kopp, Michael May, Luca Pappalardo, Nikos Pelekis, Salvatore Rinzivillo, Filippo Simini, Marios Vodas</td>
</tr>
</tbody>
</table>
Abstract

This document contains the first deliverable of WP4 (D4.1). It represents the results of the DATASIM project in WP4 during Year 2.

Copyright

This report is © DATASIM Consortium 2012. Its duplication is restricted to the personal use within the consortium and the European Commission.

www.datasim-fp7.eu
# Content

1. Introduction .................................................................................................................. 5
2. Evaluation of Spatio-temporal Microsimulation Systems .............................................. 8
   2.1 Introduction ................................................................................................................ 8
   2.2 Properties of Mobility Data ....................................................................................... 9
       2.2.1 Dimensions of Mobility Data ............................................................................. 10
       2.2.2 Observation Space ............................................................................................ 10
       2.2.3 Sampling Coverage .......................................................................................... 11
       2.2.4 Resolution ........................................................................................................ 13
       2.2.5 Missing Data .................................................................................................... 13
   2.3 Data Input and External Model Validation ............................................................... 14
       2.3.1 General Measures for Comparing Categorical and Numerical Variables .......... 14
       2.3.2 Evaluating the Distribution of Movement Positions ........................................... 16
       2.3.3 Evaluating the Distribution of Differences between Movement Positions .......... 18
       2.3.4 Evaluating Sequential Dependencies between Movement Positions ............... 19
       2.3.5 State-of-the-Art of External Model Validation .................................................. 20
   2.4 Internal Model Validation ......................................................................................... 21
       2.4.1 Validation on the Level of Model Components .................................................. 21
       2.4.2 Validation of Model Variability ........................................................................ 22
       2.4.3 Example of a Model Components Validation .................................................... 22
   2.5 Validation Use Cases ............................................................................................... 25
       2.5.1 Evaluating GPS and GSM Data in the Region of Pisa ........................................ 25
       2.5.2 Evaluating Mobile Phone Data in the Region of Lausanne ............................... 29
   2.6 Conclusion .............................................................................................................. 31
3. On Benchmarking a Trajectory DBMS .......................................................................... 32
   3.1 Introduction .............................................................................................................. 32
   3.2 Related Work .......................................................................................................... 32
   3.3 Overview of the Benchmark – Architecture ............................................................ 34
       3.3.1 Storage Model .................................................................................................. 35
       3.3.2 Partitioning Scheme ......................................................................................... 37
       3.3.3 Queries ............................................................................................................. 37
   3.4 Roadmap .................................................................................................................. 38
4. References .................................................................................................................... 40

---

Project funded by the European Community under the Information and Communication Technologies Programme - Contract ICT-FP7-270833
Tables

Table 1: Example of a confusion matrix ................................................................. 15
Table 2: Predictive performance of discrete (D) and continuous (C) decision trees .................. 24
Table 3: Comparison of average daily travel distances (in km) ........................................... 30
1 Introduction

The goal of deliverable D4.1 is to provide a novel, better and more detailed standard for the evaluation and benchmarking of agent-based spatio-temporal microsimulation systems. In this spirit we performed various kinds of validations of mobility models, simulation models and mobility sets of different sources (e.g. GPS, GSM, frequency counts). The deliverable is organized according to these topics.

Chapter 2 describes the validation of spatio-temporal microsimulation systems. The main challenge for a correct and meaningful evaluation is posed by the inherent complexity of movement data. The combination of spatial, temporal and population variables allows to derive a nearly unlimited number of mobility characteristics. State-of-the-art evaluations, however, rely on partial evaluations of only a few selected mobility characteristics. For example, comparing the simulation results with existing traffic counts (Gao et al., 2010; Horni et al., 2009; Meister et al., 2010), and thereby constraining the view on frequencies at several locations, is still the dominating approach. The increasing application range of spatio-temporal microsimulation systems requires a holistic view on validation, able to evaluate many different aspects of movement behavior. For example, many different parameterizations of a simulation model would be able to generate routes that can pass the test of comparing frequencies at a sparse set of locations, although the underlying trajectories and flow structures of its agents are very different. Therefore, tests on frequency counts alone would not be sufficient, for example, when being interested in main directions of flow.

Chapter 2 assembles available validation techniques and provides an overview on the state-of-the-art of validating spatio-temporal microsimulation systems. It collects and structures various additional aspects that have to be considered for the validation and comparison of movement data. In particular, in terms of processes, or phases, we differentiate validation of the input data, validation of the internal model as well as validation of the external model. The validation of input data is crucial because incorrect or missing values in the input data usually are propagated and emphasized through the simulation processes. Only high-quality input data leads to reliable simulation results. Internal model validation indirectly addresses the issue of how well the “inner” working of the model is “consistent with reality”. This is measured by testing how well the model is able to predict the given input data. Finally, external model validation confronts the model results with external data and determines the final quality of the model results. The chapter concludes with examples of using big data sources for the extraction and validation of movement characteristics. More details on the validation are given in the appended papers. Section 2 is a joint research effort by WP4 partners and will be published as book chapter within the “Data on Science and Simulation in Transportation Research” book.

Chapter 3 addresses a system point of view by referring to the moving object database (MOD). In the recent past, several DBMS-based solutions for storing and efficiently querying historical trajectory data were already proposed in the form MODs, (Almeida et al., 2006) (Pelekis et al, 2008) (Pelekis et al, 2011). The DATASIM project revealed the need for a MOD to support new queries that consider the semantics of a trajectory in addition to its pure spatio-temporal components. Correspondingly, benchmarking such a MOD must also address these semantic components. Chapter 3 describes a first sketch of characterizing what is at stake in this context: (1) a storage model (2) a partitioning scheme and (3) syntactic and semantic queries / operators. A concrete benchmark on this basis has not yet been implemented.

In order to summarize the results of this deliverable, and to highlight its innovation, we created the following table where each line corresponds to one its sections. In addition to a short summary of the content, the table sketches the state-of-the art and relates it to the novel contribution of the project’s results. The last column provides a sample reference to the work done.
Chapter 3 addresses a system point of view by assessing the moving object database (MOD). In the recent past, several DBMS-based solutions for storing and efficiently querying historical trajectory data were already proposed in the form MODs, (Almeida et al., 2006) (Pelekis et al, 2008) (Pelekis et al, 2011). The DATASIM project revealed the need for a MOD to support new queries that consider the semantics of a trajectory in addition to its pure spatio-temporal components. Correspondingly, benchmarking such a MOD must also address these semantic components. Chapter 3 describes a first sketch of characterizing what is at stake in this context: (1) a storage model (2) a partitioning scheme and (3) syntactic and semantic queries / operators. A concrete benchmark on this basis has not yet been implemented (which was also not stipulated in the DOW).

During year 2 of DATASIM, various validations on large-scale data domains were performed. We compared mobility characteristics derived from heterogeneous sources, developing various techniques for the extraction of similar information in different data sources. In year three we will intensify the validations and integrate them with the microsimulation model. In more detail, paper a) performs internal model validation to evaluate the stochastic error of a simulation depending on the level of geographic detail. Paper b) contains the evaluation standard presented in Section 2. Paper c) compares mobility characteristics derived from GPS data with traffic counts in Pisa, Italy while papers d) and e) compare large-scale GSM and GPS data. Finally, paper f) evaluates the performance of different mobility models.
List of Publications:


2 Evaluation of Spatio-temporal Microsimulation Systems

2.1 Introduction

Modeling individual movement behavior is a complex task and requires thorough validation throughout the modeling process. The complexity of spatio-temporal microsimulation systems originates mainly from the vast solution space of individual movements in geographic space and time. The size of the solution space is closely linked to the spatial and temporal resolution of the microsimulation, which is typically predetermined by the application. The outcome of a microsimulation is a complete mobility model for a given region, time period and population. Based on a synthetic population it provides a detailed schedule about who moves when and where using which mode of transport. This comprehensive information makes the validation of spatio-temporal microsimulation systems challenging.

In current practice the evaluation of microsimulations relies on a partial evaluation of mobility characteristics. Most often a comparison with traffic counts is performed as e.g. in (Gao et al., 2010; Horni et al., 2009; Meister et al., 2010). Traffic counts have the advantage that they can be comparably easy obtained. However, traffic counts do not contain origin-destination information and are typically available for vehicular traffic only. Thus, they cover only a small aspect of individual movement. In addition, most traffic counts are available for major roads only and are therefore not representative for the whole street network. In other words, traffic counts are a vital source for the validation of microsimulation systems but their application is limited to a subset of model characteristics.

The example of traffic counts illustrates that a new holistic validation concept for spatio-temporal microsimulation systems is needed. Only a validation which considers a broad set of mobility characteristics can ensure that the outcome of a microsimulation is a truthful reproduction of reality. The required variety of validation data to implement such a concept is just becoming available due to the advancement of information and communication technology. Thus, we are at the right moment of time to animate the discussion about a new validation standard. In addition, research on spatio-temporal data mining and analysis has made tremendous advances during the past decade. As a result, a rich set of preprocessing, feature extraction, indexing and data mining methods are available to exploit and handle spatio-temporal data.

The validation of microsimulation systems is a manifold task. In general, the validation workflow of the modeling process can be divided into three parts, namely input data validation, internal model validation and external model validation (see Figure 1). Input data validation ensures that the model is build using high-quality data. Usually, the provided input data comes from secondary data sources and has originally been collected for a different purpose. For this reason it is important to show in a first validation step that the input data is an appropriate source for the goal of the overall modeling process. Internal model validation measures how well the model can predict its input data, i.e. how reliable a predictive model will perform on (unseen) training data. A high internal quality is a prerequisite for the external quality of the model. If the input data of the model is difficult to predict, this is likely to be the case also for unknown data. In addition, as microsimulation systems typically rely on non-deterministic algorithms, the variability of the model is subject to internal model validation. Finally, during external model validation the model output is compared to either independent real world data or other model results with known high quality. This step ensures the final quality of the model results. If
the model outcome meets a given quality standard the process of modeling is finished. Otherwise, either the input data or the modeling process are subject to change and to a repeated validation, leading to a validation cycle.

In the following section we attempt to compile a comprehensive overview on the state-of-the-art of validating spatio-temporal microsimulation systems. We provide a structured review on the various aspects of movement data evaluation and give practical insights into challenging problems based on real-world examples. More detailed, Section Properties of Mobility Data recapitulates general properties of mobility data that influence the comparability of mobility data sets. Section Data Input and External Model Validation describes mobility characteristics and validation techniques that are commonly used to compare two mobility data sets while Section Internal Model Validation focuses on the estimation of internal model quality. In Section Validation Use Cases we provide real-world examples outlining the challenges of comparing mobility data sets. Finally, the Conclusion summarizes the vital points and draws a roadmap for the further advancement of evaluating spatio-temporal microsimulation systems.

2.2 Properties of Mobility Data

Due to the manifold techniques to record mobility information, mobility data sets can differ in various aspects. In order to make valid statements about the similarity or dissimilarity of two data sets it is therefore essential to have a clear understanding of the delimitation of each set with respect to its spatial, temporal and population dimension. In this section we will discuss these dimensions as well as different properties of mobility data sets. The properties include the observation space, sampling coverage and resolution. In addition, we consider missing data properties because they can influence the representativeness of a data set. Note that our grouping of mobility properties results mainly from our experience in practice. A partially differing compilation is, for example, chosen in (Andrienko et al., 2013).
2.2.1 Dimensions of Mobility Data

Mobility is inherently connected to three basic dimensions: space, time and population. Geographic space defines where movement takes place, time defines when movement takes place and the population or object dimension specifies who is moving. In this section we will give a short introduction to all three dimensions, summarized from (Andrienko at al., 2008; Körner, 2012).

Commonly, we refer to geographic space as the three-dimensional Euclidean space that co-rotates with Earth and is centered at its center of mass, i.e. the physical space we observe in everyday life. In order to specify the position of an object in physical space, spatial reference systems such as the Cartesian or Geographic reference system are used. Typically, the spatial component of mobility data is specified using two-dimensional geographic coordinates, i.e. the longitude and latitude of the moving object’s positions.

Time is a one-dimensional extent which describes the ordering and duration of events. Today Coordinated Universal Time (UTC) is used as standard temporal reference system to refer to a specific moment in time. UTC uses the Gregorian calendar to reference days. It further divides a day into hours, minutes and seconds. Due to the Earth’s rotation and its revolution around the sun, we perceive a natural structure of time into cycles. Over the year we observe the change of seasons, our working activity is typically organized by a weekly cycle, and our day / night rhythm repeats every 24 hours. Be aware that those cycles are nested and form hierarchies as e.g. year / month / day-in-month or year / week-in-year / day-in-week. Such hierarchies are especially relevant for the validation of mobility data because the data sets belong often to different time periods, and the aggregation of data into time cycles provides the only way for comparison.

The population dimension specifies of which entities the movements are observed. Entities may be animate (e.g. persons, animals) or inanimate (e.g. parcels, airplanes). As this book is placed in the area of transportation research, we tailor the following description to humans. Persons have numerous sociodemographic characteristics such as their gender, age, occupation or income. Sociodemographic characteristics are known to influence movement behavior (Curtis & Perkins, 2006; Scheiner, 2010). For example, the age determines whether we can travel independently by ourselves or whether we are allowed to drive a car while the occupation determines whether we have regular work trips. Another important variable that influences mobility is the place of living (Curtis & Perkins, 2006; Schwanen et al., 2005). For example, the trip length and preferred mode of transport varies between urban and rural areas.

In addition to the three basic dimensions that characterize movement, the movement itself can be described by physical and semantic properties. For example, movement possesses speed, acceleration, direction and turn characteristics. In addition, it has some means of transportation and is conducted with some specific activity in mind (e.g. when going to work).

Together the four dimensions provide a good way to structure properties of mobility data sets, which we will discuss next.

2.2.2 Observation Space

The observation space delimits the spatial, temporal, population and movement dimensions for which the comparison shall be made. More precisely, in the spatial dimension it restricts the region in which the movement is observed. Typically, this is a city, community or even larger administrative area. In addition, the spatial observation may be restricted to a specific type of geographic objects. For
example, instead of monitoring continuous space, we may observe movement only on the street network. Furthermore, we could observe only highways or the pedestrian area of a city.

In the temporal dimension the observation space defines the time moment, time interval or time cycle in which the movement takes place. If a data set consists of raw measurements it typically refers to a specific time period, e.g. the month January 2013. Often data sets are already aggregated to a specific time cycle, e.g. an average week. In such cases it is still important to know how the cycle relates to the greater time hierarchy. Is it, for example, an average week in January (with slippery roads in the northern hemisphere) or in June (holiday season), and which year does it represent (e.g. 1980)? Furthermore, an observation may be restricted to selected time intervals within, e.g. we may observe only the working days of a week or the daytime between 6 and 20 o’clock.

The population observation space defines the set of persons whose movements are studied. It is closely connected to the data collection process. If a survey is conducted, the population typically represents the residents of a defined area, e.g. the inhabitants of a city or country. Further attributes may be used in the selection process. For example, only persons above 18 years of age or persons who commute to work could be considered. This definition has the advantage that it is compatible with many official statistics. However, as geographic space is not a closed system, it lacks the mobility of externals visiting the area as e.g. commuters, freight carriers or tourists. In contrast, the population may comprise all persons travelling in a certain area as observed, for example, when using induction loops for traffic monitoring.

Finally, mobility data may be limited to certain mobility characteristics. For example, GPS tracking devices may be installed into the car of a person and record only vehicular movement. Induction loops have a similar effect as they monitor only motorized traffic. The studied movement may also be related to certain activities only as, for example, shopping, working or vacation.

### 2.2.3 Sampling Coverage

The sampling coverage is closely related to the observation space. As it is typically not possible to monitor the complete observation space, the data set is only a sample of the spatial, temporal and population dimension. Samples are generally characterized by their distribution and size. Both characteristics are important to know because they determine the representativeness and sampling error of a data set with respect to the observation space.

In the spatial dimension sampling processes are most visible when a decision about the placement of (a limited number of) stationary sensors has to be made. Examples of such sensors are induction loops, light barriers, cameras, WiFi or Bluetooth scanners which count the number of passing vehicles or persons. Note that those sensors are typically placed at strategic points with high traffic volume and may therefore not be representative for the observation space. Similar problems can arise if mobile sensors are used. Although each person moves in space and records data for various locations, a really large sample is needed for a complete coverage of the street network. For example, Hecker et al. (2010) analyzed a data set with 42,780 test persons of mixed GPS and CATI (Computer Assisted Telephone Interview) records of up to seven days in Germany. The trajectories covered barely 26.7% of the German street network. Andrienko et al. (2013) identify yet another bias inherent to event-based mobility data. The authors have analyzed sequences of geo-referenced Flickr data. Naturally people take pictures of interesting places, which are therefore over-represented in the data set. In order to detect abnormalities in the spatial sampling coverage a first step is to plot the data on a map. Co-location of data with certain geographic objects (e.g. highways, points of interest) or clearly delineated
sectors with / without data indicate a spatial sampling bias. However, as mobility is not equally distributed in geographic space further analyses may be required to detect irregularities.

In the temporal dimension the sampling coverage defines how frequent measurements are taken. As time is a uniform quantity in one dimension, the analysis of its distribution and sample size is easier than in the spatial dimension. Andrienko et al. (2013) distinguish between the analysis of the length and regularity of time intervals between measurements, the coverage of the observation interval as well as of relevant time cycles. If the time interval between two measurements is short enough to permit a good interpolation of an object’s position, the authors call the data quasi-continuous. Elsewise, it is considered episodic. For example, GPS data with a time interval of one second between measurements is quasi-continuous. Call detail records (CDR), which accumulate only during a user’s phone activity, are episodic. Both examples illustrate also the difference between regular and irregular measurement techniques. Regular measurements guarantee a homogeneous coverage of the temporal observation space. However, only in combination with frequent measurements, a representative temporal sample can be formed. Consider, for example, traffic counts which are observed for five minutes at noon every day. Because movement activity varies over time, it is not obvious how to relate these measurements to the movement activity of a whole day. Without an assumption about a relationship between noon and daily (or day-of-week) traffic, we would even not be able to make assumptions about the weekly or yearly variation of traffic outside of the observed five minute time intervals. Thus it is important to cover all relevant time cycles within the temporal observation space sufficiently. In order to analyze the temporal coverage of a data set, Andrienko et al. (2013) propose to plot histograms or the cumulative distribution function of either the number of measurements or the time interval between consecutive measurements for different time cycles.

Spatio-temporal microsimulations have the scope to model population movement. Therefore, the sampling coverage of input or validation data must typically be representative for some national population. As mobility data sets are often secondary data sources only, a variety of sampling biases can arise. As mentioned in (Körner et al., 2012) a first cause of sampling bias is the different affinity of people to either the companies or devices collecting mobility data. On the one hand, companies (e.g. mobile network providers) target specific customer groups. Their data collection is therefore biased towards those sociodemographic groups. On the other hand, data collection relies increasingly on the usage of mobile devices (e.g. CDR, Bluetooth). However, mobile devices are not equally distributed and used within the population, but show a clear bias towards the young generation. A second cause for a biased population sample is the uncontrolled relationship between persons and data collection devices. A person may carry multiple tracking devices (e.g. mobile phone(s), tablet) and thus be included several times. Similarly, an observed device may be shared by one or more persons (e.g. a car) and therefore represent multiple users. The assessment whether a data set contains a population bias is a complex task because most secondary mobility data sets contain only numeric identifiers due to privacy reasons. In such a case expert knowledge as well as good reasoning capabilities are required for the analysis (Andrienko et al., 2013).

Finally, a sample bias may also be introduced by selection processes in the dimension of movement characteristics. For example, Bluetooth scanners require up to ten seconds to detect all devices within their range. In consequence, Bluetooth-enabled devices that pass the sensor range with a high velocity have a smaller probability to be detected than slow-moving devices (Gurczik at al., 2012; Schmietendorf, 2011). In addition, a sample bias may be introduced during data preprocessing. For example, when cleaning GPS data, points above or below a certain velocity or trips below a given lengths may be removed as noise. An empirical detection of such biases is hard because it is not obvious for which type of bias to look in the first place. Therefore, it is important to have a good understanding of the data collection process and the performed preprocessing steps.
2.2.4 Resolution

We use the term resolution to refer to the level of aggregation or the amount of detail in a data set. In the population dimension the smallest possible measurement unit is a single person, which corresponds to the natural resolution of microsimulations. However, input and test data sets may not be that fine-grained. For example, sociodemographic characteristics are typically aggregated for larger geographic areas in order to be privacy-preserving. Similarly, movement information may be available only in aggregated form or without sociodemographic references. In some mobility data sets identifiers are routinely changed so that the association to a specific unit is lost over time.

The spatial resolution of a data set can vary between a few centimeters and several kilometers depending on the monitoring technology used to collect the data. For example, GPS data has a very high resolution while CDR data may relate to very large GSM cells in suburban areas. However, also the spatial resolution of microsimulations can vary. Some systems perform a simulation on the level of the street network as, for example, MATSim (Balmer et al., 2006) while other systems operate on the level of traffic analysis zones as, for example, FEATHERS (Bellemans et al., 2010).

In the temporal dimension the resolution of a data set corresponds to the time span of a single measurement. For example, traffic counts are typically aggregated at the level of hours while most data sets containing time stamps have a resolution of seconds or milliseconds. The temporal resolution is often confused with the temporal sampling rate. However, a data set may have a very low sampling rate (e.g. one GPS point every hour) while the temporal resolution of the measurement is very high (e.g. a timestamp of the format JJJJ-MM-DD:hh:mm:ss).

The resolution of movement characteristics can vary in a broad range. For categorical variables (e.g. type of activity, mode of transportation) the resolution depends on the employed ontology or classification system. For derived numeric movement characteristics (e.g. speed, acceleration) the resolution depends on the aggregation level of the spatial and temporal dimension.

2.2.5 Missing Data

Missing data typically originates from uncontrolled events or processes during data collection and extends over all dimensions of mobility data. For example, technical devices may be defective or the human recollection of movement incomplete. Missing data pose a problem to data evaluation for several reasons. First, the amount of missing data may be so high, that an exclusion of incomplete data records would strongly reduce the data set. Second, summary statistics may be considered for a given time interval. If the missing data is ignored (i.e. substituted with zero values) an underestimation of movement behavior will be induced. Third, a relationship between the absence of data and the mobility behavior of a person may exist (Körner, 2012). For example, people between 30 and 39 years of age show with an average of 53 kilometers per day the highest mobility while teenagers travel around 30 kilometers per day and people above 74 years travel only 16 kilometers on average (Bundesministerium für Verkehr, Bau und Stadtentwicklung, 2010). If certain characteristics of such groups relate to the intensity of missing data, for example, elder persons may be more reliable to carry a GPS device than teenagers, the pattern of missing data is not any more at random. Therefore it is important to detect and analyze a data set for missing data as e.g. proposed in (Körner, 2012; Andrienko et al., 2013; Hecker et al., 2010).
2.3 Data Input and External Model Validation

In general, the validation process of model input data does not differ from the validation process of model output data. In both cases various characteristics of the data set in question have to be assured by comparison against external data sources. Both times we are interested in mobility characteristics of a given population. Thus we have the common task to compare mobility characteristics between two mobility data sets, which is the main topic of this section. We will start by an introduction to general measures for the comparison of categorical and numerical data sets. Next, we give a systematic overview on the various mobility characteristics that can be used to describe mobile behavior. We will structure this part according to movement characteristics considering single movement positions, differences between movement positions (e.g. length and distance) and sequential dependencies between movement positions. In this way we increase the spatio-temporal complexity of the observed characteristic step by step. Finally, we will discuss the state-of-the-art of external model validation.

2.3.1 General Measures for Comparing Categorical and Numerical Variables

Due to the wide spectrum of movement characteristics, a number of different error or distance measures can be applied for the comparison of two mobility data sets. In general, each measure is based on a particular definition of error or distance, and its estimate will thus reflect the characteristic features and properties of the underlying error / distance function. Therefore, there is no absolute best measure, and the validation method has to be chosen considering the features that one wishes to evaluate and the characteristics of the data sets. In most cases it is recommendable to use several measures in order to have a comprehensive picture of the error / distance.

In this section we will introduce general measures for the comparison of (one or many) pairwise observations as well as for the comparison of distributions of numerical and / or categorical data. For a comprehensive overview we refer the reader to (Hyndman and Koehler, 2006) and (Cha, 2007).

Let \( A = (a_1, a_2, \ldots, a_n) \) and \( B = (b_1, b_2, \ldots, b_n) \) denote two data sets with pairwise observations. For all error measures we will assume that data set \( B \) contains the ground truth. If the data is categorical the error is typically specified as average of the 0-1 loss, i.e.

\[
Err = \frac{1}{n} \sum_{i=1}^{n} I(a_i, b_i)
\]

with

\[
I(a_i, b_i) = \begin{cases} 
0 & \text{if } a_i = b_i \\ 
1 & \text{else}
\end{cases}
\]

The error can be further analyzed using a confusion matrix. The rows of a confusion matrix represent the ground truth while the columns represent values of the second (predicted) data set. All coinciding data records are located in the diagonal of the table. Table 1 shows a fictitious example of a confusion matrix for evaluating a model that predicts the mode of transportation.
Table 1: Example of a confusion matrix

<table>
<thead>
<tr>
<th></th>
<th>car</th>
<th>public transport</th>
<th>bike</th>
</tr>
</thead>
<tbody>
<tr>
<td>car</td>
<td>5</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>public transport</td>
<td>2</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>bike</td>
<td>0</td>
<td>2</td>
<td>11</td>
</tr>
</tbody>
</table>

In total the ground truth data set contains 23 objects. The confusion matrix shows that of the eight actual cars the system predicted five as cars and three as public transport. Of the six public transports it predicted three correctly, two as cars and one as bike, and of the bikes it predicted eleven correctly and two as public transport. We can see from the matrix that the system in question has trouble distinguishing between cars and public transport but distinguishes well between bike and other types of the mode of transport.

For numeric data the most commonly applied error measures are mean absolute error (MAE), mean absolute percentage error (MAPE) and root mean square error (RMSE). They are defined as follows:

\[
MAE = \frac{1}{n} \sum_{i=1}^{n} |a_i - b_i|, \quad MAPE = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{a_i - b_i}{b_i} \right| \cdot 100\%, \quad RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (a_i - b_i)^2}.
\]

In addition to error functions, also various distance functions can be applied to measure the (dis)similarity between pairwise observations. Most well-known is the Minkowski distance

\[
d_M = \sqrt[p]{\sum_{i=1}^{n} |a_i - b_i|^p}
\]

which results in the Manhattan distance for \( p=1 \), the Euclidean distance for \( p=2 \) and the Chebyshev distance for \( p=\infty \). In the Minkowski family the weight of large individual errors increases with
increasing $p$, amounting to the maximum absolute pairwise difference for $p=\infty$. One way to reduce the effect of large errors is to apply, for example, the Lorentzian distance

$$d_L = \sum_{i=1}^{n} \ln(1 + |a_i - b_i|).$$

For visual inspection, numerical pairwise observations can also be depicted in a scatter plot. The closer the points are to the diagonal representing equal values, the more similar are both data sets. The linear dependence between the data sets can be quantified using e.g. Pearson’s correlation coefficient.

For the comparison of distributions, methods for categorical, discrete and continuous data exist. One way to treat categorical distributions is to view the two frequency values of each category as a pairwise observation and to apply the above described methods for numeric data. Another way of comparison is to perform a statistical test to assess how likely both data sets origin from the same statistic population. For categorical data the chi-squared test for homogeneity can be used for this purpose.

Discrete valued distributions can be treated similar to categorical distributions. However, due to their numerical nature, we can derive and compare further moments of the distribution as, for example, its mean, variance or skewness. The mean of two data sets can be compared using, for example, a two-sample t-test or Z-test whereas the variance can be compared using the F-test (under the assumption of normally distributed data).

Finally, continuous valued distributions can be compared using the Kolmogorov-Smirnov test for homogeneity. Similar to discrete valued distributions, we can compare the mean, variance or skewness of the distributions. In addition, the distribution may be discretized and treated similar to categorical distributions.

2.3.2 Evaluating the Distribution of Movement Positions

In this section we consider count-based evaluations of movement characteristics that describe the population’s whereabouts for given instances in time, space or a combination of both. We will begin with general statistics related to the population and the population’s movements. Afterwards we will consider their spatial, temporal and spatio-temporal distribution. General count-based population and mobility statistics are the:

- distribution of sociodemographic attributes,
- distribution of activities,
- distribution of trips,
- distribution of mode of transport.

Sociodemographic attributes that are closely related to a person’s mobility are, for example, age, gender, occupation or income. It is important to compare not only the distribution of each single...
attribute but also their joint probability distribution. Activities are typically the cause for mobility. For example, we travel to work, go shopping or to the cinema. An interesting mobility characteristic is therefore the distribution of trip activities, stating the percentage of trips that are made for a specific purpose. In addition, we can consider the distribution of activities within the population. Which percentage of the population performs a certain activity? How often is an activity performed on average? Körner (2012) defines a family of quantities that can be used for the evaluation of such questions. Similarly we can analyze the number of trips or home-based tours, and (the split of) the mode of transport (e.g. on foot, by bike, by car or public transport).

Next, we consider the spatial distribution of the population at a given moment or interval of time

- using a specific mode of transportation,
- travelling with a specific speed,
- performing specific activities.

The spatial distribution of the population becomes visible, for example, in the traffic volume. The traffic volume states the average or total number of passing vehicles or persons in a given time interval, e.g. an hour or a day. When derived from local sensors (e.g. induction loops, camera systems, Bluetooth), the traffic volume is available only for a selected set of locations. However, data mining techniques can be used to predict traffic frequencies for sites without sensors (May et al., 2008a, May et al., 2008b). Traffic counts can also be derived from GPS trajectory data as investigated in (Pappalardo et al., 2013). In such a case the geographic coverage of measurements increases, however, only a sample of passers-by is recorded. Traffic counts can further be differentiated according to the mode of transport or the speed class. In addition, differences between the spatial distribution of counts and other numeric characteristics (e.g. speed) can be quantified by calculating pairwise differences between the values of each location. For visualization the spatial distribution of the values or their differences can be depicted on a map using e.g. a color encoding similar to (Andrienko & Andrienko, 2010). In addition to the distribution of the population when travelling, the distribution of the population during the performance of activities can be analyzed. Similar to traffic counts, the number of visitors at train stations, shops etc. can be compared. While those visits are of short duration and show a high variability, we can also analyze the distribution of long-term activity locations as the home or work place. Depending on the spatial aggregation unit, statistics about the number of persons living or working in a given street, sector or city may be compared.

In the following we analyze the temporal distribution of the population at a certain location or in a certain area

- using a specific mode of transportation,
- travelling with a specific speed,
- at the begin / end of certain activities,
- at the begin / end of a trip.

The temporal usage of a location can be measured by continuously monitoring the time of passage of moving objects. Depending on the data source, a stream of events or an already aggregated count for a series of time intervals may be given. If one or both of the time series are discretized, the data sets have to be adapted to contain the same regular intervals of time. The temporal distribution can further
be differentiated according to a specific mode of transport, speed class or average speed. In addition to the temporal distribution of movements, the temporal distribution of activities or trips can be compared using their start or end time. For a visual comparison of temporal distributions, a temporal histogram as shown in (Andrienko & Andrienko, 2010) can be used. Note that the temporal distribution of passages may also be gained from GPS trajectory data as described in (Schreinemacher et al., 2012). Here again the problem arises that only a sample of passers-by are observed and the data may be sparse in time or space.

Finally, a joint comparison of mobility characteristics in the spatial and temporal dimension can be made. Due to the high complexity of spatio-temporal data, such a comparison typically relies on a sequential aggregation in space and time. For example, the data can be first divided into a set of time moments or intervals. For each time slice a measure comparing the spatial distribution of two data sets is computed. The computation is repeated for each time slice and then summarized in a graph by its mean or variance. Similarly, the data set can be divided by locations first. For each location a comparison of the temporal distribution is performed. Subsequently, the results are aggregated over all locations. A spatio-temporal comparison can also be aided by visualization. For example, a geographic map containing temporal mosaic diagrams as in (Andrienko & Andrienko, 2010) is able to show the full variation of some variable (or the difference between two data sets) in space and time. However, such maps are very complex because of their high information density. Visual analysis can then be aided by classifying similar situations. For example, Schreinemacher et al. (2012) cluster street segments according to the temporal distribution of passages. The street segments were then colored according to their types in order to identify spatial relationships. If a set of speed profiles is already given, each location can be assigned to its most similar profile. Spatio-temporal differences between two data sets can then be observed by visualizing the spatial distribution of the assigned profiles. Andrienko et al. (2012) performed a clustering the other way around. They determined the spatial distribution of visitors for a run of 30 minute time intervals and subsequently clustered the spatial distributions. For such a comparison of two data sets, a set of spatial distributions has to be given in advance, according to which a classification of time intervals can be made in both data sets. The results can then be compared in a time graph.

2.3.3 Evaluating the Distribution of Differences between Movement Positions

In this section we consider movement characteristics that are derived from the difference of two positions in either space or time, namely length and duration. In the most simple form, we can compile the total length (duration) of all trips performed in a given period of time, e.g. one year. The total sum can be further differentiated according to the used mode of transportation or the traversed type of street. For example, we can calculate the total number of kilometers (hours) travelled by car on highways. Instead of the total number, we can also compare the distribution or averages of the length (duration) with respect to different categories or a combination of them:

- persons,
- trips,
- locations,
- activities.
For example, the distribution or average trip length (duration) per person and day is a very characteristic information about a population’s movement behavior. Those statistics can also be calculated for single trips. Next, we can calculate the length (duration) to reach specific locations. Imagine, for example, that a shop is interested to determine the catchment area of its customers or a city council wants to estimate how far (long) incoming or outgoing commuters travel daily to reach their work location. Further, the length (duration) that people are willing to travel to perform certain activities can be compared. For activities we can additionally compute their duration. So far we have considered absolute values of movement characteristics. Another interesting feature is the relative value of the length (duration) of a route compared to the shortest possible route (in space or time). Such a comparison is especially useful if a traffic assignment step is performed during the microsimulation in order to map the movement between zonal areas to the street network. While time is a one-dimensional extent, movement takes place in three-dimensional Euclidean space. In order to measure the spatial extent (spread) of a person’s mobility, typically the radius of gyration (ROG) is used. The radius of gyration \( r_g \) is defined as

\[
r_g = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (p_i - p_{cm})^2}
\]

with \( p_{cm} = \frac{1}{n} \sum_{i=1}^{n} p_i \).

Hereby \( p_i \) represent the \( i = 1, \ldots, n \) positions recorded for a given user. In the mobility context, \( p_i \) is a two-dimensional vector describing latitude and longitude of a user’s movements.

Similar to the previous section, we can evaluate the spatial and temporal distribution of the above characteristics. When choosing the spatial dimension, the movement characteristics are typically attached to the place of living of a person or the targeted activity location. For example, from national statistics as (Bundesministerium für Verkehr, Bau und Stadtentwicklung, 2010) it is known that the average number of kilometers travelled per person and day of people that live in large cities is shorter than of people living in suburban areas. In the temporal dimension movement characteristics are typically attached to the start or end time of an activity or trip. This allows to compare, for example, whether two data sets contain the same characteristic variation of the trip length (duration) over the day or between week days and weekends.

### 2.3.4 Evaluating Sequential Dependencies between Movement Positions

In this section we describe characteristics that originate from the sequence of movement positions. In general, we can distinguish between the physical and semantic representation of movement sequences. In the first case a trajectory consists of the sequence of passed or visited geographic locations. In the second case a trajectory is a sequence of activities as, for example, “home → work → shopping → home”.

The most common representation of movement dependencies between geographic locations is an origin-destination (OD) matrix. An OD matrix states for each pair of start and end locations the number of performed trips in a given time interval. Such a matrix can also be evaluated for a specific mode of transport (e.g. for the railroad network) or for specific trip purposes (e.g. home-work trips). Comparing only the OD distribution of the start and end location of trips, however, does not guarantee that the “right” route choice has been made. Therefore, we can compare the OD distribution for movement positions within a trip. If the movement is mapped to the street network we can calculate an OD matrix for each passed street segment from those trips which pass the segment. Having derived a single
error value from the OD matrix comparison for each segment, a color-encoded map or average error value can be computed. Another way to compare local movement behavior is the calculation of flows between neighboring areas or street segments. Given a movement trajectory, its sequential information is discarded with exception to the very next step, i.e. we create a Markov chain from the data. Such flows describe, for example, turning probabilities at cross roads or hand-over information of GSM cells. If we extend such flows to contain several steps, we obtain movement rules as, for example, given that a person moves from A→ B then the probability of moving to C is 25%. The detection of movement patterns, however, is a complex task. One possibility is to find frequent sub-trajectories in the data sets as proposed by Giannotti et al (2007). Another possibility is to build a compact model of the location dependencies using e.g. Bayesian Networks (Liebig et al., 2009).

In the case of semantic trajectories, we have to find an appropriate measure to compare activity-travel sequences. First and foremost the goodness-of-fit measure has to be able to consider the different dimensions of activity-travel patterns. However, another critical point is that the measure should also be able to compare sequential information. The Sequence Alignment Method (SAM) is an appropriate measure as it complies with all of the above aspects. Introduced by Wilson (1998) in time use research, SAM has the right properties for working out comparisons between predicted and observed activity-travel patterns. SAM works by calculating a distance between the predicted and observed string of activity-travel information. Unlike the Euclidean distance, a specific distance is computed by determining the amount of effort it takes in order to make the two strings equal. To calculate the total amount of effort, a series of possible operations are performed on the strings. This way strings can be made equal by using so-called ‘identity’, ‘substitution’, ‘insertion’ and ‘deletion’ operations. The distance calculated this way is then taken as the measure of dissimilarity between the strings. However a shortcoming of this approach is that SAM can only handle one-dimensional strings. This means that SAM can analyze similarities for specific attributes of activity-travel patterns, such as activity type, mode choice, location choice, etc. but not the inter-relationships between elements of different attributes. Therefore, a multidimensional extension of the traditional SAM was developed (Joh, 1999). In this way multidimensional activity-travel patterns can be compared. However, the comparison relies only on the sequential activity information. For the inclusion of temporal information, further, sophisticated similarity measures have yet to be developed. Nevertheless, similarly to physical movement sequences, we can also extract and evaluate Markov probabilities and rules from the activity-travel sequences.

2.3.5 State-of-the-Art of External Model Validation

The validation of spatio-temporal microsimulations with external data sources depends strongly on the availability of data sources in the modeling region. Typically, three or four different aspects of the data can be evaluated. In the following we will provide a cross section of common microsimulation evaluations based on the MATSim and FEATHERS simulation systems. Most often, a validation of the traffic volume, trip length and trip duration over a given time interval or a 24 hour time cycle is performed. For example, Gao et al. (2010) evaluate the average trip duration and length per hour of day, the aggregated traffic volume for four 4 hour time periods as well as the average speed on highways for two 2 hour time intervals during peak traffic. They summarize the results using (temporal) histograms, scatter plots, relative differences, RMSE and regression analysis. Horni et al. (2009) focus in their work on the evaluation of leisure and shopping trips. They provide histograms about the average shopping trip length and duration. In addition, they provide histograms to compare the distribution of the length of shopping trips. The authors also evaluate the number of shopping activities per location and provide a map showing the locations with either the largest positive or negative differences for two configurations of the microsimulation. An evaluation of traffic counts is
performed using scatter plots as well as box plots. The latter show the absolute and relative differences over a 24 hour time cycle. Meister et al. (2010) target the evaluation of the mode of transport split. They provide histograms showing the cumulative model split for increasing trip length and duration. In addition, they visualize the deviation of the share of public transport for 5 x 5 km grid cells on a map. Similar to the previous authors, a temporal histogram and box plot are used to show the absolute and relative deviation of traffic counts in a 24 hour time interval. Kochan (2012) takes a broader view on the evaluation of mobility characteristics. The author also evaluates traffic counts and travel lengths. For traffic counts he uses Person’s correlation coefficient, and for travel lengths he calculates the total vehicle kilometers traveled per year. In addition, the author performs a comparison of trip start times and the distribution of trip origin-destination pairs. For the former he uses a temporal histogram over a 24 hour cycle. For the latter the author provides a scatter plot and calculates the coefficient of determination of a linear regression model.

The provided literature review shows that many different mobility characteristics are considered during evaluation. The validations included count-based evaluations of trips and activities. Evaluations of different modes of transport and speed were performed as well as differences between movement positions. Typically, the temporal distribution over a day was considered, and values for different locations provided to show the distribution in space. Yet, the selection also shows that a uniform evaluation standard is missing. Many characteristics are shown in temporal histograms without an explicit quantification of the error. This makes it hard to set up a validation benchmark to enable the comparison of results across different simulation platforms and data sets. In addition, most validations focus on either spatial or temporal characteristics. Combinations of both dimensions as well as the evaluation of dependencies and sequential information are greatly missing. Finally, all validations lack a holistic view on validation and concentrate only on few movement characteristics.

2.4 Internal Model Validation

The internal validation of transport demand models tests the ability of the model to predict travel behavior. It is typically performed on the level of model components and requires comparing the model predictions with information other than that used in estimating the model. If the model output and the independent data are in acceptable agreement, the model can be considered validated. As microsimulation systems typically rely on non-deterministic algorithms, the variability of a model is also subject to internal model validation. In this section we will first introduce general techniques for the validation of model components and model variability. Afterwards we discuss a practical example of internal model validation in the case of an activity-based transportation model inside the FEATHERS framework (Bellemans et al., 2010).

2.4.1 Validation on the Level of Model Components

Different techniques exist regarding internal model validation. A widely known approach is the so-called cross-validation method (Kohavi, 1998). This technique is generally used in prediction tasks when one wants to investigate how reliable the model performs in actual practice. The task hereby is to learn a model from data that is at one’s disposal as, for example, a travel survey. Such a model may be a regression model or a decision tree or any other decision support tool obtained by means of a learning algorithm. The difficulty of evaluating a predictive model is that it may possess strong prediction potency on the training data set, but might do worse in predicting unseen data. This
phenomenon is also called overfitting. Cross-validation avoids overfitting by separating the data set into two parts: one is used to train or develop the model, and the other part is used to validate the model. In ordinary cross-validation the training and validation data sets cross-over in sequential steps such that each data record has the opportunity to be in the validation set once. In practice various procedures exist for cross-validation of a model, namely: hold-out validation, k-fold cross-validation and leave-one-out cross-validation.

In hold-out validation a common way is to divide the available data set into two non-overlapping fractions: one for training and the other for validation. The test data is held out and not being used during the training phase. This way hold-out validation prevents that training data and test data overlap each other, yielding an estimation of higher accuracy for the generalization performance of the learning algorithm. A well-known drawback of this approach is that this procedure does not make use of all the data at hand and that secondly the outcomes are highly dependent on the choice for the training and validation data sets.

In k-fold cross-validation the data is first subdivided into k equal-sized data parts or so-called folds. Next, k repetitions of training and validation steps are carried out such that within each repetition another fold of the data is held-out for validating the model while the k-1 folds left are used for learning. An advantage of this approach is its accurate performance estimation. However, the overlapping of training sets between repetitions is a drawback.

The last validation technique discussed here, leave-one-out cross-validation (LOOCV) is a special application of the traditional k-fold cross-validation where k equals the total number of records in the data set. In each iteration step all data records except one single record are used for training and the model is tested afterwards based on that single record. The model accuracy estimation accomplished by means of LOOCV is almost unbiased.

2.4.2 Validation of Model Variability

Activity-based models of travel demand using a micro-simulation approach inevitably include stochastic error that is caused by the statistical distributions of random components. Indeed, for making choices based on decision trees the transport demand system needs to make choices by means of randomly picking out a choice alternative based on the probability distribution in the decision tree nodes. As a result, running a traffic micro-simulation model several times with the same inputs will obtain different outputs. Analysis of the impacts on the model outputs thereby is one of the vital steps in the model development and validation. In order to take the variation of outputs in each model run into account, a common approach is to run the model multiple times and to use the average value of the results. The concept of confidence interval can then be applied with the purpose of determining the required minimum number of model runs to ensure at least a certain percentile of zones in the concerned study area reach stability i.e., with a certain level of confidence that the obtained average value of each of these zones can only vary within an acceptable interval. However, how many runs are really needed in order to reach stability depends strongly on the kind of activity-based transport demand model under concern.

2.4.3 Example of a Model Components Validation

In this section we present a concrete example of a hold-out validation for an activity-based transport demand model implemented in the FEATHERS framework (Bellemans et al., 2010). We first provide a
short description of the FEATHERS framework and the activity-based transport model inside FEATHERS in order to sketch the validation context. Subsequently we discuss the results of the model components validation.

The FEATHERS framework is a versatile system that facilitates the development and maintenance of activity-based models for transport demand. For this purpose FEATHERS provides all necessary tools to develop and maintain activity-based models in a particular study area. Currently, the FEATHERS framework incorporates the core of the ALBATROSS Activity-Based scheduler (Arentze et al., 2005). This scheduler assumes a sequential decision process consisting of 26 decision trees that intends to simulate the way individuals build daily schedules. The output of the model consists of predicted activity schedules. They describe for a given day which activities are conducted, at what time (start time), for how long (duration), where (location) and, if travelling is involved, the transport mode used and the chaining of trips. The activity-based model inside FEATHERS uses a CHAID-based decision tree induction method (Kass, 1980) to derive decision trees from activity diary data. The following example validation therefore describes the quality of the resulting trained decision trees for each step in the decision process model. Most decision trees involve a choice between discrete alternatives, for example, the transport mode. However, activity duration and activity start time decisions are modeled as a continuous choice and therefore a continuous decision tree is constructed for each of these kinds of choices.

We tested the predictive performance of the decision tree models using a hold-out sample, i.e. only a subset of the data was used during training to build the model. For each decision step, a random sample of 70% of the cases (training set) was used to build the decision trees. The other subset of 30% of the cases (test set) was presented as unseen data to the models for the validation. The accuracy of each discrete choice decision tree was calculated as percentage of correct predictions while the accuracy of continuous valued decision trees was evaluated using MAPE (see Section General Measures for Comparing Categorical and Numerical Variables).

In order to make a better judgment about the predictive ability of the discrete choice decision trees, we additionally calculated a null-model for each decision tree. A null-model is defined as a model that randomly selects a choice alternative. The performance of a null-model can simply be derived by dividing 100 by the number of choice alternatives. By comparing the null-models with the respective decision trees the relative performance of each tree can be determined, i.e. it becomes possible to see whether or not the decision trees are vigorous enough to score better then the null-models.

As stated before the activity-based model inside of FEATHERS consists of 26 decision trees. However, for illustration purposes, we show the validation of only a selection of decision trees. The predictive performance of the selected decision trees on the training and validation sets are presented in Table 2. As can be seen all selected trees perform better then the null-models where a random choice alternative is being selected. Indeed, when looking, for example, at the first tree, which predicts whether or not a work activity is included in an activity-travel schedule, it can clearly be seen that its performance is much better than of its equivalent null-model. The null-model indicates that it would correctly predict choices in 50% of the cases while the accuracy on the test set is 77.8% indicating that the tree correctly predicts choices in almost 78% of the cases. Based on Table 2 it can also be concluded that the degree of overfitting for the selected decision trees, i.e. the difference between the training and the validation set is low. Therefore it can be underlined that the transferability of the model, with regard to the selected trees, to a new set of cases is satisfactory. Keeping the first decision tree as an illustration again, it shows that the training and the test set accuracy differ only by about 0.1% meaning that the decision tree that was estimated with the training set clearly performed well on the unseen test set.
For all continuous valued decision trees in *Table 2* the MAPE, determined on the validation and test set, is shown. As can be seen, the MAPE for each decision tree is approximately the same for the training and validation set, implying that the decision trees perform well in case of unseen data cases. However, values differ much across all trees. This is caused by the fact that the nature of the different choices to be determined is very diverse. For example, the duration of work episodes tend to be rather stable as opposed to the duration of fixed episodes (e.g. bring / get activity). Nevertheless, overall the continuous decision trees perform quite well.

<table>
<thead>
<tr>
<th>Choice</th>
<th>Tree type</th>
<th>Nr of choice alternatives</th>
<th>Null model (%)</th>
<th>CMA Training set (%)</th>
<th>CMA Test set (%)</th>
<th>MAPE Training set (%)</th>
<th>MAPE Test set (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inclusion of work episode</td>
<td>D</td>
<td>2</td>
<td>50.0</td>
<td>77.9</td>
<td>77.8</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Total duration of work episodes</td>
<td>C</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>27.9</td>
<td>26.4</td>
<td>-</td>
</tr>
<tr>
<td>Timing of work episodes</td>
<td>C</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>11.8</td>
<td>12.9</td>
<td>-</td>
</tr>
<tr>
<td>Work location, in/out home</td>
<td>D</td>
<td>2</td>
<td>50.0</td>
<td>63.1</td>
<td>61.4</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Transport mode work episodes</td>
<td>D</td>
<td>4</td>
<td>25.0</td>
<td>65.3</td>
<td>62.1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Inclusion of fixed episode</td>
<td>D</td>
<td>2</td>
<td>50.0</td>
<td>87.2</td>
<td>86.8</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Duration of fixed episodes</td>
<td>C</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>140.1</td>
<td>154.0</td>
<td>-</td>
</tr>
<tr>
<td>Timing of fixed episodes</td>
<td>C</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>33.3</td>
<td>34.6</td>
<td>-</td>
</tr>
<tr>
<td>Inclusion of flexible episode</td>
<td>D</td>
<td>2</td>
<td>50.0</td>
<td>79.6</td>
<td>78.8</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Duration of flexible episode</td>
<td>D</td>
<td>3</td>
<td>33.3</td>
<td>41.8</td>
<td>39.5</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Timing of flexible episode</td>
<td>D</td>
<td>6</td>
<td>16.6</td>
<td>49.5</td>
<td>47.4</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Location, same as previous</td>
<td>D</td>
<td>3</td>
<td>33.3</td>
<td>60.0</td>
<td>59.1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Location, distance-size class</td>
<td>D</td>
<td>25</td>
<td>4.0</td>
<td>8.9</td>
<td>7.7</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Transport mode non-work episodes</td>
<td>D</td>
<td>4</td>
<td>25.0</td>
<td>52.7</td>
<td>49.8</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

*Table 2: Predictive performance of discrete (D) and continuous (C) decision trees*

The example above demonstrated the application of the hold-out validation technique in case of an activity-based transport model inside FEATHERS, however other validation techniques discussed previously may be considered as well.

---

1 A dash (-) in the table indicates that the respective error measure is not applicable
2.5 Validation Use Cases

In today’s world mobility surveys are still the main source to gather data as input for mobility models. Unfortunately the implementation of a survey does have some significant disadvantages. It is usually time-consuming and connected with costs. For this reason surveys are often restricted in the number of participants, space or time. In consequence, it is often hard to build a comprehensive mobility model that offers detailed information for a large geographical area for a longer period of time given the input data.

In contrast to surveys, more and more big data is available for analysis. In most cases the data is not tailored to the use in mobility models. Yet, it offers a high potential to describe mobility of daily life. Two of the most common big data sources are GPS and GSM. For this reason we will explore the evaluation of mobility characteristics based on both big data sources in this section. This is done by two examples. In the first example we compare a GPS data set in a central region of Italy with a GSM data set of a European country. This evaluation is done by an indirect cross-data set validation were we compare derived statistics from two distinct data sets portraying the same phenomenon. In the second example both data sets are collected in parallel in the greater area of Lausanne, Switzerland.

2.5.1 Evaluating GPS and GSM Data in the Region of Pisa

Radius of Gyration

In this section we present a cross-data set validation using GPS and GSM data. The employed GPS data set contains information of approximately 9.8 million different car travels from 159,000 cars with on-board GPS devices. The GSM data was collected by a European mobile phone carrier for billing and operational purposes. With respect to the dimensions of mobility data discussed in Section Properties of Mobility Data, both data sets differ in several aspects. Regarding the observation space, the GPS data refers to trips performed during one month (May 2011) in an area corresponding to a region in central Italy (a 250 km x 250 km square). In contrast, the mobile phone data covers an entire European country and a period of observation of six months. Moreover, with respect to the population, the car travel data set represents a 2% sample of the overall population of cars in Italy, while the mobile phone data set covers users of a major European operator (100,000 users). The Resolution of the GPS data is higher than of the GSM data, providing very detailed information about the spatio-temporal position of users, with an average sampling rate of a few seconds. Conversely, information provided by the GSM data set is not very accurate in terms of space and time because an individual may be anywhere within a tower’s reception area, which can span up to tens of square miles. Since call patterns are bursty, for most of the time we do not know the actual position of the user.

Despite its low resolution, mobile phone data is very appropriate to study general mobility. It usually includes all possible means of transportation. In contrast, GPS data may refer only to a particular kind of mobility. For example, our study data set contains only car traces because the GPS devices were installed into the cars. The fact that one data set contains aspects missing in the other data set makes the two types of data suitable for an external validation of patterns emerging from human mobility behavior. We do not expect to observe the exact same behavior in both data sets, but the same tendencies and laws behind the movement patterns.

The example of external validation we discuss in this section has been analyzed in detail by Pappalardo et al. (2013). The authors investigate whether known general mobility patterns found in GSM data by González et al. (2008) also apply to car travel. Based on the GPS data set described
earlier, they computed main mobility measures used in literature, such as the radius of gyration (see Section Evaluating the Distribution of Differences between Movement Positions) and compared it to the GSM data set. The distribution of the radius of gyration across the GPS data set resulted in a power law with an exponential cutoff: \( P(r_g) \sim (r_g + r_o)^{\beta} \exp(-r_g/\tau) \) with \( r_o = 5.54 \), \( \beta = 1.13 \) and \( \tau = 39.76 \) (see Figure 2 left). Though the parameters differed from the earlier estimated parameters \( r_o = 5.8 \), \( \beta = 1.65 \) and \( \tau = 350 \) (see Figure 2 right) by González et al. (2008), the type of the curve agrees with the previously found results. The results confirm that the vast majority of individuals tend to travel within small distances, whereas some of them carry out very long journeys. The results further strengthen the insight that a huge heterogeneity exists in the characteristic travel distance of people. Moreover, the variability seems to be independent from the spatial observation space (a region with GPS data vs. a whole country with GSM data) and temporal observation space (one month for GPS, six months for GSM). As we see in the GPS plot of Figure 2 (left), there is a difference between the predicted behavior and the observed behavior for people with a radius of less than \( \sim 5 \) km. This is presumably due to the sampling coverage, since people tend to cover small distances by foot, bike, or bus, resulting in a low probability to find such travels in the car data set. This is a phenomenon we do not observe in GSM data, since the data covers all types of travel.

Another interesting characteristic of individual mobility is the most frequent location \( L_1 \), i.e. the zone where a person can be located with highest probability when not moving, most likely the home or work place. Estimating \( L_1 \) in GSM data is rather immediate: it is the tower from which the user performs the highest number of calls. Working with GPS car traces is more complex because the data does not provide explicit information about the visited locations of a user but only of the used parking sites. In order to solve the problem Pappalardo et al. (2013) applied the Bisecting K-means clustering algorithm (Tan et al., 2005) on the sets of origin and destination points of the sub-trajectories. The most frequent location \( L_1 \) does not necessarily coincide with the center of mass \( p_{cm} \) of movement positions, which is

![Figure 2: Distribution of the radius of gyration; left: computed on the GPS data set (source: Pappalardo et al., 2013); right: computed on the GSM data set (source: González et al., 2008)](image)
used in the calculation of the radius of gyration. Pappalardo et al. (2013) therefore calculated the distance $d(p_{cm}, L_1)$ between both locations and related it to the previously extracted radius of gyration. The results are depicted in Figure 3 for the GPS data (left) as well as the GSM data (right). In both data sets the distance tends to grow with the radius of gyration. However, the GSM data shows an interesting “trail” emerging for higher radius of gyration that requires further investigation. The strong correlation between the two variables is interesting and presumably due to the systematic nature of human motion. Indeed, if a person travels arbitrarily in any direction from and to the same preferential location, the distance between the center of mass and the most frequent location will tend to zero, and the radius of gyration will have no relation with it. On the contrary, since each vehicle follows systematic travels among few preferred places, the center of mass is pulled by these trips towards the mean point of the frequent locations. Therefore, the more a vehicle travels away from its $L_1$, the more the center of mass tends to be distant from the most frequent location. This outcome in both data sets suggests that the center of mass is not adequate to describe a realistic barycenter of individual mobility, especially in the case of users with a large radius of gyration.

![Figure 3: Correlation between the radius of gyration $r_g$ and the distance between the center of mass and the most frequent location of users $d(p_{cm}, L_1)$; left: computed on GSM data set (source: Pappalardo et al., 2013); right: computed for GSM data set](image)

In the last example the comparison between GPS and GSM is made only visually. In a future analysis the values will also be compared numerically. In order to do so we need first to properly choose the size of the bin used to sample users, second to choose an appropriate function to describe the correlation and third to fit the data to it.
Traffic Counts

Consider a data sample representing movements of individuals such as, for example, the GPS car travel data set introduced above. A comparison with the ground truth, i.e., the movements of cars of the whole population within the urban area, would be very useful both for validation and for traffic prediction purposes. Although in general it is very difficult to obtain a whole population describing a particular phenomenon, nowadays many sensing technologies are available that provide streets-based traffic counts.

In (Pappalardo et al., 2013) the authors use a data set containing logs from Variable Message Panels (VMP) to assess the generality of analytical results obtained from the GPS data set. VMPs are devices situated in the entry gates of the city of Pisa with the purpose of counting all entering vehicles. The information provided by VMPs has a coarse population resolution because it does not tell which cars pass the devices but only how many vehicles pass through the gates. In terms of observation space, the two data sets cover the same time period (although the VMPs cover a larger temporal extent). The VMPs are constrained on specific positions over the road network. Thus, exploiting the spatial precision of GPS data, all vehicle trajectories were intersected with the VMP locations. As for the temporal accuracy, the VMPs provide an aggregated count of vehicles on an hourly basis. The determined VMP passages in the GPS data were therefore aggregated hour by hour. Figure 4 shows the hourly frequency counts of the VMP and GPS data of one gate in a time graph. It can be clearly seen that there is a good match between the curves, which essentially differ for a scaling factor. This is due to a different sampling coverage of the two data sets: VMPs are able to count any vehicle passing, whereas the GPS data set contains only traces of a subset of all vehicles.

In order to perform a scaling between the VMP and GPS traffic counts, the authors applied a discrete wavelet transform (DWT). A DWT is a mathematical tool that projects a time series onto a collection of orthonormal basis functions and produces a set of coefficients, capturing information from the time series at different frequencies and distinct times. From the coefficients the authors build a model to scale the traffic counts obtained from the GPS data sample to the full population. Figure 5 shows the real VMP series along with the scaled GPS signal and the measured relative error at a selected VMP location. The error is low when the GPS traffic is high. During the night hours the relative error tends to grow since there are too few circulating GPS vehicles, but the absolute error is still negligible. However, for a traffic manager it is crucial to have a precise estimation during the rush hours in order...
to design ad hoc intervention options to avoid congestions, a situation for which our reconstruction provides a very high precision.

![Graph of GPS signal comparison](image)

**Figure 5**: Comparison between real and scaled GPS signal at a single gate in Pisa

### 2.5.2 Evaluating Mobile Phone Data in the Region of Lausanne

In our second example we explore GPS and GSM data in the area around Lausanne, Switzerland. The data was collected during the Lausanne Data Collection Campaign, which was carried out by the Nokia Research Center (NRC). During this campaign the NRC equipped about 200 people living around Lake Geneva with smartphones (Nokia N95). Each smartphone, equipped with multiple sensors, collected a wide range of data generated by each participant. All data was continuously collected for about one year. In 2011 a data set of 38 persons was released to the research community (http://research.nokia.com/page/12000).

The data set itself consists of both social and geographic data. This includes information about the GPS Position and the GSM-Cell-ID in which a mobile phone activity, like calls and SMS, took place. With respect to the dimensions of mobility data discussed in Section *Properties of Mobility Data*, the data set covers a large temporal observation space (one year), but is limited in its spatial observation space (area around Lausanne). Especially the population sample (38 participants) covers only a very small part of the Lausanne population. The resolution of the GPS data is, like in the first example, higher than that of the GSM data. GSM only produces a record in the data set if the phone of a participant is interacting with the mobile network of the operator. GPS, on the other side, is producing a record every second if the quality of the satellite connection allows it. Another point to mention is that
in this second example the GPS data is not limited to a particular kind of mobility (e.g. cars) since the GPS data origins from smart phones, which were carried around by the participants all day long.

This combination of GPS and GSM data makes the data set very interesting, because it gives the rare opportunity to compare GSM data with ground truth information about the mobility of a person. For this reason we evaluated the usability of mobile phone data for the extraction of mobility quantities. In the following we present an excerpt of our results published in (Schulz et al., 2012). In particular we were interested in measurement quantities that help to enrich and improve existing mobility models. Two of the analyzed quantities are presented in the following.

**Travel distance**

In the first analysis we compared the travel distance based on the GPS and GSM data sets. The GPS data include the distance between any two consecutive GPS points excluding those points inside of a stop as well as the travel distance considering only the centroid coordinates of stop locations. Against it, the GSM data set contains the average daily travel distance between consecutive GSM activities predicated on the estimated GSM-cell centroids. See the table below:

<table>
<thead>
<tr>
<th>GPS Sequence</th>
<th>GPS Stops</th>
<th>GSM Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>39.10</td>
<td>19.22</td>
<td>18.56</td>
</tr>
</tbody>
</table>

*Table 3: Comparison of average daily travel distances (in km)*

The results in Table 3 show that the distance calculated from GSM activity data is only a half of the travel distance measured by GPS. This is a sign that GSM Data cannot characterize the real daily travel. On the other hand, if the travel distance is reduced to distance calculated from GPS stops, both measurements are similar. This leads us to the second analysis, which is about the identification of frequent activity locations.

**Activity Locations**

In our second analysis we took a closer look at frequent stop locations. These are defined as locations where people stay over a longer period of time to do some activities, like work, school or sports. For this reason we can also call those stops activity locations. After identifying those locations from the GPS data, we conducted two analyses. In the first one we determined the proportion of GSM activities that take place within typical activity locations. In the second one we analyzed the number of stop locations that can potentially be detected through GSM activity data. In both cases more than 2/3 respectively 50% of all frequent activity locations could be detected. In sum this means that GSM activity data is a good source to identify and analyze activity locations. On the other hand it also means that GSM activities mostly tell us about where people stay, not where they move, which is consistent with our analysis about the travel distance.
2.6 Conclusion

The increasing interest in spatio-temporal microsimulation systems as well as the diversity of available spatio-temporal data sets call for a new evaluation standard for microsimulation systems. On the one hand, the standard has to direct researchers and practitioners towards a holistic view on movement validation. On the other hand, it has to broaden the scope of validation methods to seize the potential of new mobility data sources. This section compiles a comprehensive overview on the state-of-the-art of validating spatio-temporal microsimulation systems by providing a systematic overview about properties of movement data, mobility characteristics, commonly used similarity measures and validation schemes. A cross section of the state-of-the-art shows that an explicit quantification of the external model error is often missing, which restrains the comparison of results across different simulation platforms and data sets. In addition, most evaluations are limited to a small set of mobility characteristics, focusing typically on either spatial or temporal characteristics. Furthermore, comparisons are typically performed on the level of persons, trips, activities or movement sequences. The validation of movement characteristics based on group patterns (e.g. convergence, moving clusters), however, has not been considered in the literature so far. The major reason for this practice is the complexity of movement data and the limited availability of external validation data. The former problem requires a tight interaction between the mobility mining research community and the transportation research community in order to turn currently complex analysis methods into standard tools that are generally available. The latter problem is likely to decrease with the increasing availability of big data sources. However, big data sets bring their own challenges as the data may cover only a part of the observation space, differ in its temporal resolution or be not representative in all aspects. Our validation examples using real-world application data show that differing data properties often hinder a direct comparison, and additional research efforts have to be invested into the development of data harmonization techniques. In summary, the validation of spatio-temporal microsimulation systems is a complex task and vital research area which holds many challenging research questions for future work.
3 On Benchmarking a Trajectory DBMS

3.1 Introduction

A MOD system combines many different aspects of Mobility under one common framework that has to be mature enough to support operations such as data loading, indexing, basic and advanced querying, and data mining (e.g. clustering) efficiently. Also, the experience we gained from the research project DATASIM recently, revealed the need for a MOD to support new queries that consider the semantics of a trajectory in addition to its spatio-temporal components. The only unbiased way to prove a system’s maturity and measure its performance characteristics is to test it using a suitable benchmark.

We consider a MOD benchmark to consist of three basic dimensions: (1) storage model (2) partitioning scheme and (3) queries / operators which cover all the aspects mentioned in the previous paragraph. Each dimension can have multiple alternative implementations, therefore we can, for example, compare the performance of a range query on two different storage schemes. Measurements can either be based on the total execution time of an operation or the total space requirements that a storage option has on disk. Also, an important contribution is that, to our knowledge, this is the first attempt to introduce semantic queries to a benchmark.

After reviewing the related work in MOD benchmarking in the next section, we then provide a theoretical description of our benchmark in section 3.3 and conclude with summarizing the current status and describing the next steps in section 3.4.

3.2 Related Work

In this section, we present the current status of other benchmarks and we attempt to identify the main characteristics of each approach.

The state of the art framework in the literature for benchmarking MOD engines is the BerlinMOD (Düntgen et al., 2009). The paper focuses on moving object data generators and two sets of queries for benchmarking MODs, namely range and k-NN queries.

First the authors present the spatio-temporal data types of the SECONDO system that are used to implement the generator and the set of queries. The main data type is the moving point and it is modeled using a “sliced representation” where an instance of this type is expressed by a set of units (i.e. spatio-temporal segments). Besides, the data types the authors also present the operators that are used to implement the set of queries.

The benchmark scenario uses synthetic vehicle positions generated using a graph representation of the road network of Berlin, Germany. The vehicle’s owner is assigned a HomeNode, a WorkNode, and a Neighbourhood and these nodes are used to generate the trips of the owner between his home and working place. The scenario assumes that trips from home to work and back to home happen from Monday to Friday between 8 a.m. and 4 p.m. There is a chance that additional trips might occur in the spare time and in the weekends which will always end at home.
To generate each trip a start and a destination node are selected and then a shortest path algorithm is invoked between them. By following the path while taking into consideration the maximum allowed speed of each road segment that the driver always respects (in fact he will accelerate in order to reach that maximum speed every time). Also, the driver will sometimes stop because of traffic, obstacles, etc. For the trips generated during the spare time the destinations are usually chosen from the neighborhood of the vehicle owner.

The authors propose two ways to store the data, one that keeps the whole history of an object in one row (OBA) and another that each trip of the object is in its own row (TBA). Some relations are common to both approaches, such as: (1) Nodes, that contains all road network nodes (2) QueryPoints, randomly chosen nodes from Nodes to be used in the queries (3) QueryRegions, random polygons (4) QueryInstants, timestamps uniformly distributed within the observation period (5) QueryPeriods, time periods where the starting timestamp is generated as in QueryInstants and the duration (in days) is sampled from a Gaussian distribution (6) QueryLicences, contains the license plate numbers of the vehicles. The moving object data are stored in different ways in the two approaches. In OBA the movement data are stored in a relation named dataScar along with other information like the car type and model. In TBA the relation dataMcar contains the car type and model information and the relation dataMtrip contains the trips.

In order to categorize the set of the range queries, the authors identify five query properties (1) Object identity (known/unknown) (2) Dimension (standard/temporal/spatial/spatio-temporal) (3) Query interval (point/range/unbounded) (4) Condition type (single object/object relations) (5) Aggregation (aggregation/no aggregation). By evaluating all the possible combinations of the previous query properties the authors eliminate those that are meaningless and then proceed with the presentation and explanation of the final queries.

The k-NN queries are only theoretically investigated as a possible extension to the benchmark in the future. The problems identified are that there is little work done on k-NN queries for historical trajectory data and that there is no standard way to formulate those queries in SQL. The authors identify three query properties (1) Condition type (k-NN/r-NN/a-NN) (2) Query object type (static/moving) (3) Candidate type (static/moving).

In (Theodoridis, 2003), the authors propose a specific schema for storing the trajectories of a hypothetical scenario (no real or synthetic data) that involves shoppers moving around an area by walking or using different transport modes. Based on that scenario the authors categorize a collection of queries and discuss their performance on different kinds of spatio-temporal index access mechanisms that exist in the literature.

The database schema stores the movement history of a shopper in one row using the moving point data type, like the OBA paradigm in (Düntgen et al., 2009). There are other relations for storing auxiliary data such as roads (polylines), buildings/shops (polygons), and shop offers (temporal periods).

The benchmark queries include (1) loading and update operations (insert and update records and index entries) (2) range/distance/k-NN/topological queries on stationary reference objects (3) distance/k-NN queries on moving reference objects (4) join queries (5) unary operators.

The authors also present a taxonomy of spatio-temporal indices based on whether each structure indexes past or current locations thus supports asking past/current/future queries.

Both (Düntgen at al., 2009) and (Theodoridis 2003), propose using a generator to create synthetic datasets to be used for the benchmark.
Other, well-known benchmarks like (Stonebraker et al., 1993) and (Myllymaki and Kaufman, 2003) are usually limited to the spatial dimension of trajectory data with little or no focus on including the temporal dimension in their data model and queries.

3.3 Overview of the Benchmark – Architecture

In general, a benchmark’s goal is to test the system’s performance on as many “base case studies” as possible in order to understand and measure its strengths and limitations. The main idea of the proposed benchmark is the concept of dimensions which aims to identify and separate the most influential factors in a benchmark. In our approach we identify three dimensions:

- The “storage model” dimension is the way we choose to store the data in a relational system. Essentially, it specifies the structure of the table’s record.
- The “partitioning scheme” dimension is the way we partition (or not) our spatio-temporal data. The options here could vary significantly since any attribute of the dataset can be chosen as the partitioning key. In our approach we narrow this to three alternatives all of which are based on range partitioning.
- The “query” dimension is a common and useful question in the spatio-temporal domain that is posed as an SQL query (e.g. range or k-NN query) to the MOD system and could be implemented in multiple ways w.r.t. the available storage models. The goal is to measure the execution time of every implementation of every query.

Figure 6 shows a visual depiction of the dimensions and the operations of the benchmark.
3.3.1 Storage Model

The storage model dimension of the benchmark is probably the most important one and it is also the one that is affected the most by the data types and the access methods the MOD system supports. Assuming that the data is hosted in one table (either real or virtual in the case of partitioning), we identify two basic options in this dimension the segment-based and the trajectory-based storage models. In the first option each row of the table is a segment of the trajectory formed by two consecutive sampled time stamped positions whilst in the second option each row contains the whole trajectory of a moving object. The segment-based approach is, of course, the most fine grained because a segment is the most basic part of a trajectory that represents movement.

The weaknesses of each approach are becoming obvious when we want to build appropriate spatio-temporal indexes, usually a variation of the 3D R-tree (Theodoridis et al., 1996), for each one of them. In the “trajectory per row” solution the indexed minimum bounding boxes of the trajectories are expected to contain excessive dead space. On the other hand, the index on the “segment per row” solution contains a much higher number of nodes (because every segment occupies a row in the table) so its growing size is expected to affect its performance.

These shortcomings led to a hybrid third option where each row contains a sub-trajectory that preferably introduces the smallest possible dead space. This sub-trajectory could have been created based on semantic information, i.e. each STOP / MOVE episode is considered a sub-trajectory (see Figure 7), or alternatively by applying a segmentation method on the trajectory to identify sharp changes.

DATA INPUT

This benchmark focuses on trajectory data that might or might not be semantically enriched (Parent et al., 2013). It is important that the data input formats are unbiased w.r.t. the loading procedure and allow the update of existing data in the database since in a real-world cases the objects will continuously generate new trajectories as they move.

We propose three CSV-based raw input formats that are generic enough to fulfil the previous recommendations:

- Basic-CSV file «ObjectID, TrajectoryID, Timestamp, Lon, Lat»: This is the simplest format in which every row contains a sampled position of the object’s movement (see Figure 8).
  - Note that a dataset in which the movement history of an object is not segmented into trajectories can be represented with this input format simply by setting TrajectoryID always to 1.
- SubTraj-CSV file «ObjectID, TrajectoryID, SubTrajectoryID, Timestamp, Lon, Lat»: This format is the same as Basic-CSV, only it is more refined since every position is assigned to a sub-trajectory.
- Semantic-CSV file «ObjectID, TrajectoryID, SubTrajectoryID, Timestamp, Lon, Lat, Episode, Tags»: Each episode contains multiple contiguous positions of a trajectory thus it is also a sub-trajectory (Figure 7). Episode can either be “STOP” or “MOVE” and Tags can take any text value the user wants (see Figure 9).

Because data usually arrive continuously in batches (i.e. a dataset will be composed of multiple files) and in order to be closer to reality all formats assume that the rows are sorted ascending according to time and that a trajectory reconstruction phase has happened beforehand.
For a benchmark to be complete both a synthetic and a real dataset should be used if possible. Unfortunately, finding a real dataset that is large enough in size to accommodate the benchmark’s scalability experiments is not always possible. The lack of suitable real datasets has led to the development of many spatio-temporal data generators (Düntgen et al, 2009), (Theodoridis and Nascimento, 2000) (Brinkhoff, 2002). However, proposing a new generator is out of the context of this section, although in the experimental evaluation of our benchmark we explain how we used a data generator to increase the size of a small semantically enriched real dataset.

```
... 237397000,1,2008-12-31 19:29:30,23.5429166584333,38.0261883206083 0261883206083
239157000,1,2008-12-31 19:29:30,25.3790933261087,36.440536655699
239190000,1,2008-12-31 19:29:30,23.5537499918725,37.9476999874264
239203000,1,2008-12-31 19:29:30,25.2334783258433,36.6963183219695
239212300,1,2008-12-31 19:29:30,23.627833251772,37.9346499874
...
```

*Figure 8: Sample of Basic-CSV file*
LOADING

The loading operation takes a batch in a data input format and inserts new trajectories or updates them if they already exist, i.e. trajectory refresh. For each call of the loader on a batch the total time of execution is kept in order to see how the loader behaves as new data is inserted in the table and what is the overhead that the spatio-temporal index introduces. Once the loading procedures complete for two or more storage models then by comparing the results on space efficiency of each storage model useful conclusions could emerge.

3.3.2 Partitioning Scheme

The majority of the spatio-temporal indexes proposed in the literature are based on partitioning time, space, or both (Choi et al., 2004) (Ni and Ravishankar, 2007), Cude-Mauroux et al., 2010). Therefore, it is very interesting to benchmark how each kind of partitioning performs in conjunction with the different storage models when a query is executed.

In cases where the volume of the data is very big, which in the case of trajectory data are continuously updated with new positions, then storing everything in a single table could eventually affect performance of the queries even if an index is present.

3.3.3 Queries

In this section we identify the types of queries that are most common in the spatio-temporal domain and for each query type we provide the corresponding expression in natural language. The actual SQL expression of a query is different in each MOD system, therefore, they will not be discussed in detail for now.
BASIC QUERIES

Some of the most heavily used queries in trajectory databases are the timeslice, range, and k-NN queries. The three query types are expressed in natural language below:

- Q1: “Find the positions of all objects at a specific timestamp”.
- Q2: “Find the movement of objects inside a specific area and at a specific timestamp”.
- Q3: “Find the object that approached closest to a place and also find when it happened”.

JOIN QUERIES

This type of queries is very interesting because it involves operations between two sets. By benchmarking this query it is possible to see how the system performs under a heavy operation such as joins. In Q4 there is a self-join query:

- Q4: “Find pairs of objects that were located closer than a specific distance at a short period of time; for each pair, find their minimum distance of one object from the other”.

OLAP FASHION QUERIES

OLAP queries have the potential to indicate how the MOD system performs on demanding and complex procedures such as the calculation of origin-destination matrices. The next query can be such an operation:

- Q5: “Perform equi-sized homogeneous partitioning in space and in time (e.g. a 10x10 grid in space and a 1-day interval in time) and count the number of objects per cell”.

MINING QUERIES

Mining trajectory data usually means that some kind of clustering must take place. These operations are quite common because by analyzing the result the user can easily identify typical routes or outliers in the movement of the objects. However, they are resource intensive and for that they are supported by some kind of index. Practically, a trajectory clustering algorithm must be applied in way similar to Q6:

- Q6: “Apply a clustering method on the trajectories using a similarity measure between trajectories”.

SEMANTIC QUERIES

So far semantic queries are not investigated in a benchmark framework for MODs. Semantic queries involve pattern matching operations in conjunction with spatio-temporal predicates. The queries below present a few common semantic queries:

- Q7: “Find the objects that stopped inside a specific area and within a period of time”.
- Q8: “Find the objects that begin their trips from home then go to work and do not return to home in a specific time period”.

3.4 Roadmap

Hermes MOD is currently in a mature state where it can support a wide range of storage models and queries. Hermes has been extended to support the querying of semantically enriched trajectory data in
addition of course to the pure spatio-temporal querying support (i.e. range, k-NN, join, OLAP, mining etc.). The benchmark has been designed to test how Hermes performance scales when executing queries on big volumes of moving object data. The next steps in this effort are towards applying the benchmark on Hermes using semantically enriched datasets and to compare the results with those of a simulated MOD on top of PostGIS. Preliminary tests have shown that Hermes performs better than PostGIS in MOD-related operations.
4 References


