2017

Evaluation of Geospatial Features for Forecasting Parking Occupancy Using Social Media Data

Johannes Riedel

University of Rhode Island, johannesriedel1@gmx.net

EVALUATION OF GEOSPATIAL FEATURES FOR FORECASTING PARKING OCCUPANCY USING SOCIAL MEDIA DATA

BY

JOHANNES RIEDEL

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE
IN
INDUSTRIAL AND SYSTEMS ENGINEERING

UNIVERSITY OF RHODE ISLAND

2017

MASTER OF SCIENCE THESIS

OF

JOHANNES RIEDEL

APPROVED:

Thesis Committee:

Major Professor Jyh-Hone Wang

Valerie Maier-Speredelozzi

Christopher Hunter

Nasser H. Zawia
DEAN OF THE GRADUATE SCHOOL

UNIVERSITY OF RHODE ISLAND

2017

ABSTRACT

Urbanization and growing individual mobility are globally active trends that intensify the needs for transportation in cities. In this context, parking space has become a scarce resource. Drivers searching for open parking spots cause about one third of the total traffic in urban areas. This creates significant fuel consumption, greenhouse gas emissions and time loss. Intelligent Transportation Systems with particular focus on parking are a promising approach to overcome the information asymmetry and lead drivers directly to available parking spots. This requires highly accurate occupancy data for parking areas on a geographically extended scale.

The ultimate goal of this thesis is to improve the modeling of parking occupancy by extraction of meaningful features from raw data in social media. The research focus is set to points of interest and public events in urban areas. First, robust methodologies are developed for the acquisition and benchmarking of large-scale social media data. This includes exploratory data analysis and testing of Facebook as a leading platform against alternative online data sources. Here, a multi-stage approach for the identification of duplicates in heterogeneous data sources is applied. Secondly, a diverse set of feature extraction methodologies is developed that integrates a variety of secondary data sources and findings in the literature. This comprises the adjustment of online popularity attributes for social media objects based on external data and the extraction of parking-related attributes based on text mining. Additionally, historical parking events from Floating Car Data are cross-referenced to thematic similarities among objects and adequate feature sets are derived. This includes the category-specific transformation of historical parking patterns into characteristic time- and object-dependent features. Also, text-based topic modeling using Latent Dirichlet Allocation is applied on social media data to extract thematic object similarities as probabilistic input features for parking demand modeling. In the final evaluation phase, ground truth occupancy data for a selection of off- and on-street locations is used to compare machine learning models trained with varying input feature sets. A baseline and extended set are compared while the latter includes extracted social media features. These models account for the prediction of parking occupancy over different timeframes. Random forest learning machines that include social media features are found to outperform the tested baseline models for both off- and on-street parking demand modeling. Particularly event topic probabilities and category-specific parking events on an hourly basis are identified to be valuable.

Contents

Chapter

Page

Page

Page

List of Figures

Figure **Page**

Figure **Page**

List of Tables

CHAPTER 1

Introduction

1.1 Urban relevancy of parking space availability

Urbanization and growing individual mobility create new challenges for today's city infrastructures. Especially in emerging economies, increasing urban population and individual wealth lead to a new level of mobility needs. This expresses in terms of rapidly growing passenger cars numbers due to the fact that they are perceived as status symbols in many cultures. The number of vehicles in China, for instance, almost tripled between 2007 (59 Million) and 2015 (172 Million) [1]. Also in developed economies such as Germany, the total number of passenger cars still undergoes slight increases, even though the population is rather declining [2]. An extension of existing infrastructure is required but typically is found infeasible due to budget regulations or space limitations [3].

In this context, parking space can be considered a precious resource in the urban environment. Empirical studies have shown that roughly one third of the total city traffic is caused by drivers searching for available parking [4]. This consumes resources, causes noise and increases air pollution. In fact, the quality of life in urban hotspot areas decreases remarkably with intensified traffic and shortage of parking space. Another result is the increase of illegally parked vehicles that cause macroeconomic costs being estimated to $80 Million per year alone in the city of Barcelona, Spain [5]. The individual time for drivers to find urban parking generally varies among cities but typically ranges between 3.5 to 14 min as has been stated by Shoup 2006 [6]. Therefore, parking must be considered an important factor when planning mobility and deciding on a certain travel mode [7].

1.2 AIPARK parking information platform

Artificial Intelligence Based Parking (AIPARK) is an *Intelligent Transportation System* (ITS) that provides comprehensive information related to the parking situation in cities. The system evolved from research activities at TU Braunschweig starting in late 2015 and provides data for currently more than 60 Million parking spots in Germany. AIPARK's main purpose is to guide drivers to available parking space near their travel destination. An overview on the elements of AIPARK is provided in Figure 1.

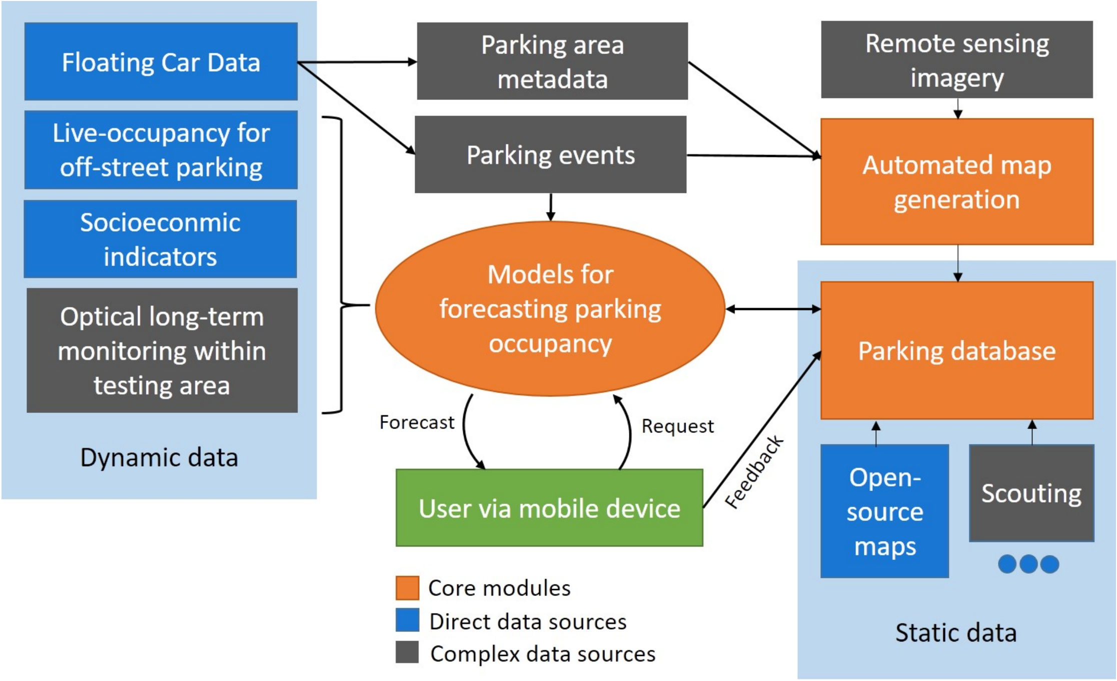


Figure 1. Elements of the AIPARK platform

AIPARK is implemented as a scalable platform that includes modules for data acquisition, processing, modeling and user interaction. Moreover, a comprehensive database of static information is provided that contains the location of parking areas and relevant metadata such as opening hours, pricing or parking restrictions. This database was initially based on open-source map data derived from numerous contributions of volunteers. These sources undergo crowd-based review processes and can be used within the AIPARK system without major adjustments. Thus, they are referred to as 'direct data sources' in Figure 1. At a later point in time, modules for the automated generation of parking maps are used to refine the existing information and extend the coverage of the parking database. A core technology is the analysis of remote sensing imagery, such as satellite or aerial images for the purpose of map generation. In the course of this step, geographic locations are identified where on-street parking is possible. Significant research and development efforts are involved in extracting valuable information from the raw imagery. Therefore, this type of data is referred to as 'complex'.

The second mapping process is focused on the generation of parking area metadata based on the analysis of *Floating Car Data* (FCD). The latter refers to positional information generated by GPS or mobile devices that are placed within vehicles. By accumulating data from a number of sample vehicles over time, conclusions on traffic flow and driver behavior can be drawn [8]. AIPARK uses the concept of *parking events*, an approach that focuses on identifying when drivers leave or successfully find a parking spot based on FCD analysis. Also, negative parking events are considered, denoting the unsuccessful search of available parking spots indicated by certain driving patterns. Minor corrections of the parking database are conducted based on local *scouting*, the manual on-site collection of data using specific mobile applications.

Another core module of AIPARK are machine learning models that provide occupancy information for urban parking areas in the static database. Diverse dynamic data sources are acquired, prepared and transformed into valuable input features for model generation. Besides parking events from FCD, occupancy information for off-street parking facilities and sensor-monitored on-street spots are considered. Also, diverse contextual factors are taken into consideration, such as socioeconomic indicators. The latter comprises area-specific statistical data for the factors car ownership, income level, business activity and age distribution. As a subproject, long-term optical observation of parking areas is conducted for strategically important parking spots in urban areas. This is based on the automated extraction of occupancy information from camera footage that monitors parking areas over several weeks. To maintain this complex data source, costs and timewise efforts are comparably high. This is why long-term optical observation is primarily conducted within the city of Braunschweig that serves as a testing area for the AIPARK platform.

1.3 Research design

1.3.1 Primary research questions addressed

Social media platforms are one of the most rapidly growing sources of multifaceted online data of our time. Users interact and share their personal data in different explicit and implicit ways on a regular basis. Social media is expected to reflect trends, opinions and behavior in society and on a personal level. Significant research has addressed the question of how the available information can be used to derive new insights in the field of mass mobility. However, the aspect of parking demand has not been sufficiently addressed despite its relevancy due to globally increasing car usage.

This thesis aims at improving the understanding of geographically referenced social media data and its relevancy as a data source for modeling urban parking demand. A special focus is set to two different types of objects: Points of interest (POIs) and public events. The research design approaches the domain from a theoretical, literature-based perspective, as well as from a practical, implementation-related side. This guarantees that relevant findings are also evaluated with regards to their real-life implications and scalable feasibility.

1.3.2 General methodological overview

Addressing the research questions previously introduced, the methodological framework developed in this thesis is subdivided in four subsequent phases. They are presented in Figure 2. In the first phase, different online data sources for POIs and events are preliminarily evaluated and leading social media and alternative platforms are selected for further investigation. Subsequently, large-scale data is acquired from these sources using their publicly accessible application programming interfaces (APIs). This phase is referred to as *Data Acquisition* (Chapter 4). As the AIPARK project is primarily active in Germany, scalable data collection procedures are developed to acquire the target information on a nationwide scale. Dense data availability is a basic requirement of drivers using ITS. In fact, also potentially relevant social media data as input for parking demand modeling must be available with significant coverage. This phase is crucial to prove the general technical feasibility.

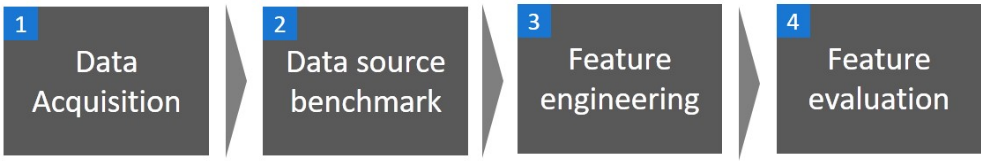


Figure 2. General phases of the developed research methodology

As part of the phase *Data source benchmark* (Chapter 5), the acquired datasets are compared to specify the inherent value interconnected with their individual integration into the forecasting system. This helps understand the benefit of social media against alternative data sources. An important step in this phase is the identification of duplicate objects among datasets using text mining and supervised learning algorithms to get reliable estimates on the scope of the examined databases.

The phase *Feature engineering* (Chapter 6) describes the information retrieval procedures used to transform raw social media data into a variety of potentially valuable sets of input features. The applied techniques are primarily based on data fusion, integrating findings from the literature and text mining.

In the last phase, *Feature evaluation* (Chapter 7), a testing procedure is developed to evaluate the previously extracted features. This covers their implications on the occupancy of both off- and on-street parking. This comprises the training of multiple machine learning models while comparing the prediction performance of a baseline featureset to an extended version that includes the extracted social media features. The benefit of specific extraction procedures is also evaluated.

CHAPTER 2

Contextual background

This chapter provides an introduction to established technical approaches used for solving the parking problem in urban environments. Also, basic concepts in data mining are briefly described and the characteristics of social media as a source for geospatial information are investigated. Moreover, certain mobility indicators for Germany are introduced as it is the geographic focus area of this study. This is expected to be valuable for contextual understanding of the modal split indicators observed and to make the derived findings more comparable to other focus areas in future studies.

2.1 Popular solutions to the urban parking problem

There are several alternatives that focus on improving the availability of parking in cities. The simplest option is to have governmental actions focus on increasing the infrastructural capacity to an extent where no parking shortage occurs. Typically, this is seen as highly unrealistic for most scenarios based on the necessity of significant public investments and land use. Also, the inherent improvement potential is very limited with regards to dense historic city centers or highly frequented areas that would be affected by extensive construction work. For this reason, the usage efficiency of existing infrastructure must be increased.

2.1.1 Parking guidance systems

Information systems that guide drivers to available parking spots improve the infrastructural utilization and are currently widely spread. These approaches fundamentally rely on the public availability of detailed parking information with regards to the destination area. By different means, this knowledge is shared among drivers using public display boards or smartphone applications. Recent approaches also consider the distribution of parking information using vehicular ad hoc networks [9]. In consequence, the time consuming search for open parking spots is minimized.

On one hand, based on discrete event simulations for a single parking lot, Surpris, Liu,and Vincenzi [10] only identified insignificant time gains due to the introduction of a parking information system. This is interpreted as a result of the limited scope of the study. On the other hand, Caicedo et al. 2006 [11] stated reductions of required searching times of up to one third when trying fo find parking in multilevel garages. Generally, driver acceptance for IT-supported applications that deliver parking information is very high [12]. These systems are found to trigger several positive effects for urban life such as reduced traffic congestion and decreased searching time [13]. Guidance systems are most likely to be used by drivers who are unfamiliar with the destination area [14].

2.1.2 Stationary sensors

As of today, existing systems cover mostly only parking garages or other paid areas. These are generally referred to as *off-street* parking. Here, occupancy data can be easily acquired as digital entrance barriers and sensors are widely distributed. Parking operators are primarily interested in collecting occupancy data for management insights and often also share this information with the public. The vast majority of parking spots in cities, however, is present in the form of *on-street* parking. In this context, occupancy information is not as easily accessible and must be generated using specific sensors that have to be primarily financed with local governmental budgets. Many different stationary systems for parking surveillance have been developed. Popular concepts cover radar sensors installed on street lights [15], camera-based surveillance using large-scale image processing [16] or magnetic field sensors integrated into the ground [17]. All of these approaches correspond with high expenses for installation and maintenance. Moreover, each of the systems can only cover a very limited amount of parking spots. One of the largest pilot projects involving parking sensors was the *SFpark* project. 6,000 systems were installed at an estimated cost of approximately USD 1.5 million. However, the project only covered slightly more than two percent of the estimated city's 281,000 on-street parking spots [18]. In fact, full-scale coverage with stationary parking sensors is very unlikely due to limited public budgets. Therefore, stationary sensor systems cannot be considered a general solution for the urban parking problem.

2.1.3 Crowdsensing

An alternative to stationary sensors is the implementation of *crowdsensing* systems that dynamically collect data and automatically extract certain geospatial features. Cruising vehicles have been used as mobile ultrasonic sensor nodes that generate dynamic maps of vehicles parked on-street. Mathur et al. 2010 [18] used these sensors to detect signal patterns that relate to parked cars on the street sides while the measurement vehicle was normally driving. Evaluating the collected data, an overall information accuracy of more than 90 percent was achieved. Moreover, the effect of attaching sensors to a population of taxicabs was simulated and significant savings were predicted when using crowdsensing instead of stationary parking sensors.

Several approaches exist that focus on smartphone data for obtaining occupancy-related information. Rinne and Törmä 2014 [19] combined *geofencing* and *activity recognition* to detect when drivers are located on a designated parking area and try to find an open spot. For instance, if designated parking lots are highly occupied in reality, users tend to leave without parking and continue searching for an alternative. If spots are sufficiently available, this is indicated by successful parking events. The system suffers from the fact that every status change of a parking lot requires at least one user that cannot find an open spot immediately. Also, the procedure is not applicable to small parking lots that drivers can easily overview without entering. In this case, no trace of unsuccessful searching for parking is found in the generated movement data.

Other researchers examined the potential of magnetic field sensors in smartphones to detect nearby vehicles [20] [21]. As cars are typically built from a significant share of ferromagnetic material, they cause magnetic perturbations that can be measured as deflections of the geomagnetic field. The detection principle has successfully been used within stationary sensors [17]. Mobile systems face limitations due to dynamically changing environmental variables and low sensitivity of the measurement devices. Even though the preliminary results are promising, no fully functional system for identification of open parking spots was yet published. Besides, mobile payment records as another source of information generated by smartphones, also have been used to derive parking occupancy information [22].

2.2 Introduction to data mining

This section provides a summarized overview on the state of the art in data mining and related sub-disciplines. It introduces typical workflows in the field and creates a simplified schema with regards to the variety of different concepts. For detailed explanation of the machine learning algorithms used in the course of this thesis, it is referred to more in-depth literature.

2.2.1 Cross Industry Standard Process for Data Mining

The *Cross Industry Standard Process for Data Mining* (CRISP-DM) is the standard reference model of the data mining field. It was introduced in the mid

1990s by a consortium of industrial companies which, at the time, were leading in applying data analysis techniques. Nowadays, it serves as one of the basic concepts for the data mining field. It consists of six phases that are relevant for both commercial and scientific use cases, and is independent from specific platforms and tools used. An overview is provided in Figure 3. The framework's main purpose is the facilitation of communication among analysts, customers and other stakeholders. CRISP-DM helps structuring data analysis projects and provides general guidance. Each top-level phase consists of diverse lower-level tasks, checklists and recommendations [23].

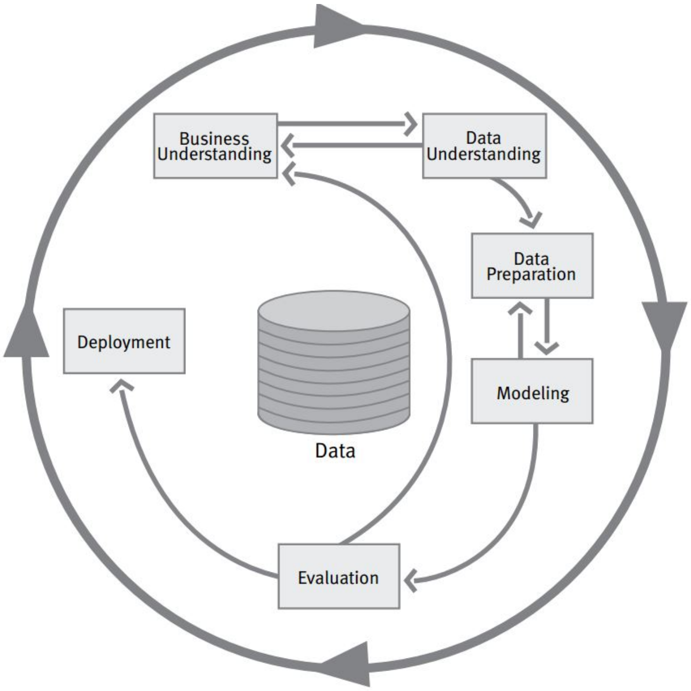


Figure 3. CRISP-DM top-level phases [24]

The first phase *business understanding* is focused on defining the objectives of the analysis project and deriving specific data mining goals and success criteria. This includes assessing the given situation regarding resources, requirements, preliminary assumptions and potential risks involved. Also, the cost-benefit ratio of the data mining project must be considered. Subsequently, during *data* *understanding*, raw data is collected and its characteristics are analyzed. This leads to a verification that the available data quality is sufficient for further usage. Afterwards, in the course of *data preparation*, low-quality data is corrected or removed and meaningful data is selected to be used in downstream procedures. Among others, this comprises merging of heterogeneous information, formatting and construction of new attributes from existing ones. During the *modeling* phase, adequate techniques are chosen to fit the specific requirements of the available data. Non-linear relationships among attributes, for example, can only be modeled with adequate learning algorithms that are not limited to linear functions. State of the art implementations of these algorithms typically offer a broad range of possible parametrization that affects the model performance. To assess the model quality, the available data is divided into subsets for training and testing. The model quality is determined by its respective accuracy on the test set which contains unseen values. Using revised configurations, modeling is a highly iterative process. During the subsequent *evaluation* phase, the model outcome is compared to the originally defined success criteria and the entire data analysis process is reviewed. In case there are potential improvements, the process is restarted in any of the preceding phases to correct errors or extend the scope of actions. If the evaluation indicates successful completion of the project, *deployment* of the obtained results into productive systems can follow. However, in the context of CRISP-DM, this phase mainly comprises planning and monitoring activities of the deployment [24].

2.2.2 General concepts in machine learning

In machine learning, it is generally distinguished between *supervised* and *unsupervised* analysis problems. In the supervised case, models are generated to predict a given target attribute based on a variety of input features. *Classification* is one major subgroup that focuses on predicting discrete target data based on a set of labeled training samples. *Regression* tasks denote settings where the target variable is continuous. For unsupervised learning, there is no corresponding target attribute for the given feature vectors. Here, the ultimate goal may be exploratory data analysis or grouping of somehow similar data. This is referred to as *clustering.* Evaluation of the generated models is based on separating the available data in a set for training and testing. As the latter remains unseen, it represents an adequate basis for examining the achieved generalization of the model. *Cross-validation* is a common process of iteratively separating the available data into changing training and testing subsets to evaluate the model accuracy while avoiding overfitting [25].

2.2.3 General procedure for data source selection and feature integration

The scope of potentially important attributes that have an influence on the urban parking situation is comparably large. To give examples, socioeconomic indicators, public holidays or weather conditions are identified to be relevant in literature. Thus, machine learning models intending to reflect the local parking situation have to incorporate these influence factors.

Figure 4 visualizes the process of integrating new attributes into predictive machine learning models. As a first step, multiple hypothesis have to be drawn regarding the relevance of certain feature groups. If findings in the literature, specific domain knowledge or common sense indicate the potential importance of a certain factor, data sources reflecting this information are identified, accessed and evaluated. With regards to parking, common sense for attribute candidate selection is a reasonable practice as it represents an everyday problem. Chosen data sources must contain the valuable attributes with geographical references. Depending on the scope, focus and character of the respective databases, this can be an open-source project or a proprietary resource with nationwide or more narrow coverage.

In practice, the data formats used and the quality of information provided can greatly vary. Thus, the interface-dependent procedures that are necessary for data aggregation have to be tested for feasibility and scalability. As the ultimate objective of AIPARK is superior coverage, the underlying data sources - individually or in combination - have to provide adequate coverage. Data sources that are not accessible with reasonable efforts or lack of quality or coverage cannot be used as a basis for parking occupancy modeling.

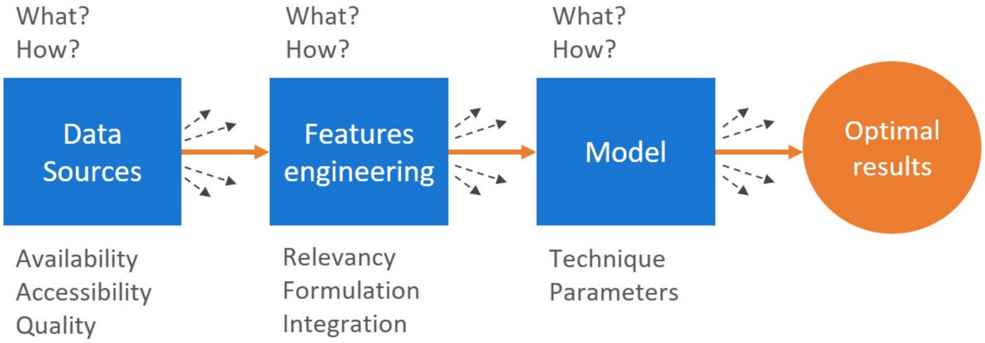


Figure 4. Key limitations and decision points for modeling

As soon as the potentially relevant sources are selected, there are plentiful options to extract and select valuable features from the available attributes. Information retrieval procedures are applied on the directly accessible attributes to obtain meaningful features for the subsequent model generation. Regarding the modeling itself, there is a multitude of machine learning algorithms that can be applied with nearly unlimited options for parametrization. All in all, the identification and selection of data sources, the formulation of features from the available data, as well as the training, optimization and selection of different modeling approaches is an extremely time-consuming procedure. The optimal results with regards to the achieved predictive accuracy of the generated model can only be realized if all of the preceding phases of the actual modeling are aligned and well conducted.

2.2.4 General concepts in feature engineering

Formulation and selection of relevant input parameters, typically referred to as *features*, is the most labor-intensive element of building machine-learning models. The success of data analysis depends significantly on the input feature vectors [26]. Thus, preprocessing of data is considered to be the most important step in deploying data mining applications [27]. As an example, the winning contribution of the popular 2010 KDD cup data mining competition credited data preparation as their key to success [28].

The term *feature engineering* comprises both the *construction* and *selection* of valuable attributes. Feature construction increases the dimensionality of the problem. Based on a set of raw information, different strategies can be applied to obtain higher-order attributes. This process is typically manual and demands certain knowledge of the problem sphere. One frequently applied technique is the decomposition of categorical features. For example, if the attribute values are sorted into three classes whereas one of these represents the value 'unknown', the latter can be instead included as a separate binary feature that gives an indication on the availability of sufficient data. This avoids that a lack of data is considered a separate class. Moreover, continuous variables can be separated into bins that comprise a certain value range to obtain a transformation into categorical attributes. This can improve the understanding of data. Also, changing of units may have positive effects [29]. From a theoretical perspective, the number of attributes constructed can be infinite. Automated feature construction supports the growth of data dimensionality [30]. However, in reality, computational resources limit the feasible model complexity. Also, to obtain adequate model accuracy, the required amount of samples grows exponentially. This phenomenon is often referred to as the *curse of dimensionality* [31]. Especially distance-based models, for example nearest neighbor classifiers, perform badly in hyperspace [32]. Using a smaller number of attributes to train the model facilitates data visualization, storage, handling and ultimately leads to better model performance [26].

One approach to reduce the present dimensionality is the combination of features into more meaningful representations. *Principal Component Analysis* (PCA) constructs a linear combination of multiple attributes while maximizing the retained variance in the dataset. *Multidimensional Scaling* (MDS), on the other hand, focuses on maximizing the distance between data points in the lower-dimensional space. Besides these classical techniques, a large number of non-linear approaches exist. It is referred to Van der Maaten et al. 2008 [33] for further information and a comparative study.

After having completed the construction of potentially relevant variables, *feature selection* techniques can be applied. It is distinguished between *filter* and *wrapper* methods. Filters are advantageous regarding their consumption of computational resources as they evaluate the importance of attributes independently from a chosen modeling technique. Wrappers, on the other hand, choose an optimal feature subset based on the relative performance of multiple models being trained with different features. The selected learning algorithm is similar for all compared feature subsets and no model-specific effect on the prediction performance is to be noted. As a third category, *embedded methods* for feature selection are considered. These are model-specific and integrate the evaluation of attributes into the model training process [26].

Both filter and wrapper approaches are based on a variety of search strategies. A popular selection criterion within the filter category are correlation indices between input features and objective variables. Attributes with stronger correlation are considered to be generally more relevant. However, selecting only the most important features typically is not optimal as it promotes redundancy [26]. An alternative concept focuses on *single variable classifiers*, which comprises training of multiple models only with a single input parameter to be evaluated. The accuracy of the obtained predictions is used as selection criterion for the input parameters. One major disadvantage of this approach is that the observed model performance highly depends on the interaction of dataset and model. Thus, different modeling approaches used for the same input parameter can lead to divergent results [34]. A third alternative is the feature selection in accordance with *information-theoretic criteria*.

When applying an *exhaustive search* strategy for wrapper methods, all potential attribute subsets are evaluated separately. Especially for large datasets with extremely high dimensionality, this is not feasible due to limitations of computability within reasonable time frames. Thus, *heuristic search* strategies are applied. For example, *forward selection* begins with a single attribute and adds relevant features step-by-step. Alternatively, in *backward elimination*, all attributes can be considered for the initial feature set, being followed by stepwise attribute removal. To determine the ranking of features to be included or removed, typically information-theoretic criteria are used [35].

Guyon and Elisseeff 2003 [26] summarize the strategic procedure of feature construction and selection with a ten-point checklist. It is focused on the practical implementation of the respective techniques in the field and visualized in Figure 5. The order of the phases presented does not follow the proposed workflow of CRISP-DM and may occur in a recurrent or iterative manner. Thus, the sequence shall not be understood in a chronological manner.

After the data is explored and sufficiently understood, certain relevant features can be created 'ad hoc' if sufficient domain knowledge is available. Subsequently, if

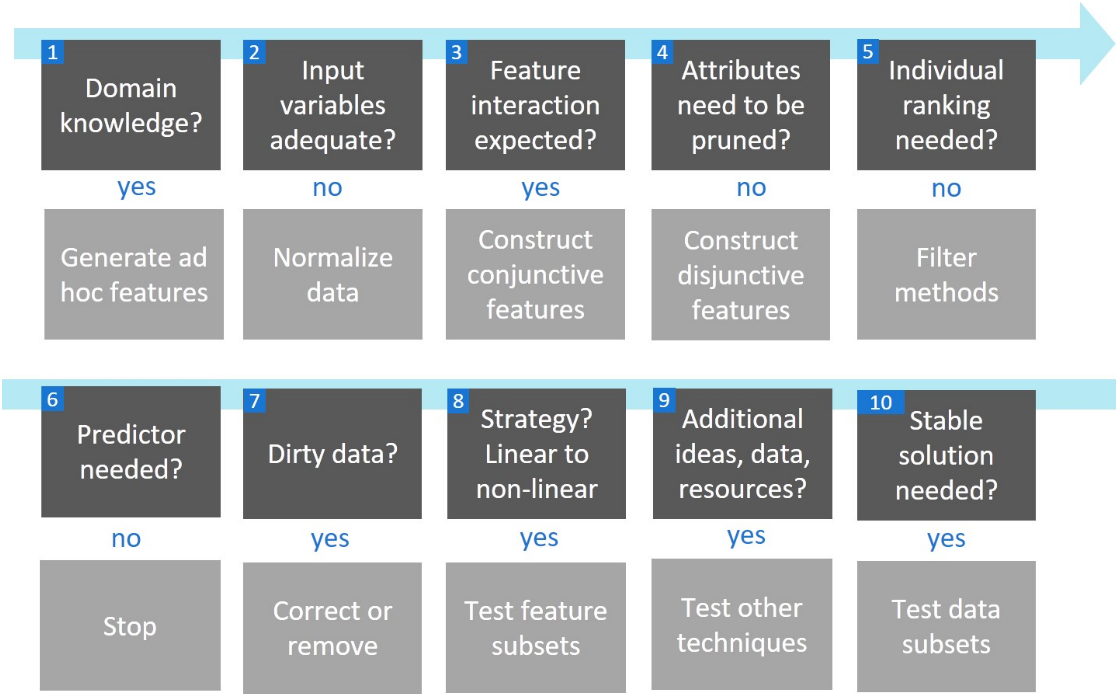


Figure 5. General checklist for feature engineering; based on [26]

the data is skewed or needs other adjustment, it is normalized to minimize the introduced bias for subsequent steps. If interdependencies of variables are expected, it is beneficial to construct abstract features from the original data using different mathematical operations. These are called *conjunctive features* and represent higher-order interactions. If the computational resources are limited, summarized basic features can be constructed to reduce the problem dimensionality. These are referred to as *disjunctive features* and may exemplarily denote weighted sums. If the influence of single variables is to be understood, filter methods can be applied. In case the main project purpose is the feature exploration, the analysis can be stopped at this point. For most cases, a generalized model predicting the objective attribute represents the ultimate analysis goal. In terms of preparation, *dirty data*, especially missing or wrong values, need to be corrected or removed. Values that are randomly missing and data where there is an underlying pattern must be distinguished. Missing values at random can at least partly be corrected by interpolation. Missing data based on an specific effect is usually hard to reconstruct. Here, case-specific solutions must be found [36].

The next phase focuses on training various models with different subsets of features. If the underlying relationship between feature and objective attribute is known to be either linear or non-linear, adequate learning machines can be directly chosen. If this information is not available, Guyon and Elisseeff 2003 [26] suggest starting with simple modeling techniques requiring only reasonable computational resources for training. Thus, linear models are to be chosen first and shall be followed by the implementation of non-linear ones while constantly comparing their performances. The authors suggest using forward selection as strategy for feature subset generation. After this phase is completed and sufficient time and resources are still available, other techniques with higher resource consumption can be examined. This may include backward selection or embedded feature selection methods. To maximize the stability of the generated predictor, it shall be tested on different subsets of the data, for example using cross-validation.

Heaton 2016 [27] examined the relationship of different feature construction approaches and the performance of modeling techniques using synthesized datasets. It was found that artificial neural networks (ANNs) and support vector machines (SVMs) perform well on features that are calculated as differences and ratios of basic attributes. For random forests and gradient boosting machines, rather aggregated and count-based features are found useful. This is seen as an important reason why superior performance is frequently observed for ensemble learners that rely on individual models from both classes.

2.2.5 General concepts in text mining

Text mining is a subfield of data mining that focuses on the extraction of information from textual data. It is part of the research field *natural language* *processing* and uses specific methodologies that apply to unstructured data. The latter makes data cleansing and feature preparation highly complex. Natural language processing must deal with ambiguous expressions and highly depends on background knowledge for the analyzed data [37]. The following paragraphs provide a short summary of common techniques in the text analysis domain organized in a chronological order within a typical workflow.

The first step in conducting a text mining project is usually the acquisition of a text corpus - a collection of documents from a specific source or thematic distribution. All further analysis is based on the information distributed in the corpus. Single documents within this collection are typically represented as sparse and high dimensional matrices. Each word is used as a feature that may occur a certain number of times within a given document. This leads to computationally highly expensive analysis operations.

A basic representation of text used for machine learning is called *bag-of-words*. Here, single words are treated as a set of occurrences while their order and grammar are not taken into account. For documents where the order of words carries valuable information, *n-grams* can be used as features. Here, a set of n subsequent words is treated as a single unit to account for spatial relationships among words. Text strings are subdivided into bags-of-words using *tokenizers* that are based on a syntactic ruleset. Simple tokenizers separate entities at whitespace characters while more complex algorithms may also account for known expressions, for example including punctuations. If it is beneficial for the problem domain, tokenization can also be applied to separate entire sentences within a corpus.

Instead of representing word occurrences in a document using binary attributes, the respective frequency of terms can be used to assign term weights. *Term frequency* (tf) is defined as the number of term occurrences in a document w

divided by the total number of words in that document n. This approach considers all terms equally relevant as features. However, many words may only have limited discriminating power with regards to the conducted analysis task. For example, a corpus consisting of documents related to biological descriptions of flowers may show high term frequencies for the terms 'blossom'. In this case, this feature is rather not helpful to distinguish between documents and may cause confusion of the learning machines used downstream. To overcome this problem, the *inverse document frequency* (idf) metric is introduced. It measures the term importance by reducing the feature weight for frequent terms and scaling up for rare terms. It is calculated as the logarithm of the number of documents N divided by the *document frequency* (df), the number of documents that contain the feature. Finally, a combined weighting scheme for each term in each document is generated by multiplying both metrics. This approach is called *tf-idf*. It assigns high weights to terms that are found multiple times within a small portion of documents, reflecting a higher discriminating power of these. Tf-idf is calculated as presented in Equation 1 while t refers to a specific term and d to a specific document in the corpus.

$$\text{tf-idf}_{t,d} = \text{tf}_{t,d} * \text{idf}_t = \frac{w_t}{n_d} * \log \frac{N}{\text{df}_t} \tag{1}$$

As sparsity is a common characteristic of feature spaces from textual data, dimensionality reduction is frequently applied to reduce computational costs and improve the model quality. One common method is the removal of *stop words*, frequently used words that do not contain valuable information. Being generally applicable for entire languages, stop words may be articles or prepositions. More specifically for a certain problem domain, also customized terms might be irrelevant. It has to be noted that a phase-based search often depends on terms that may generally be considered to be stop words. The n-gram ’flights to Berlin’, for example, crucially changes its meaning if the stop word ’to’ is left out. In fact, there must be a problem-specific decision whether or not to remove stop words. Other methods are focused on summarizing similar words into one feature. A common technique is called *stemming* and aims at reducing inflectional word forms by removing the word suffixes. It is based on a heuristic process that does not always obtain the actual word stems as results. *Lemmatization*, on the other hand, uses lexical information and a morphological analysis to return a human-readable base form of the word. The latter is referred to as *lemma*. The set of terms ’am, are, is’, for example, can be projected on the verb infinitive ’be’ [38].

Common unsupervised learning methods for text data are clustering and *topic modeling*. Both problems are closely related while clustering produces a hard segmentation of different groups and topic modeling results in soft membership probabilities for a given document. Each of the identified topics in the corpus represents a probability distribution of word features. If it is known that documents in a given corpus can contain a variety of themes, hard assignment of documents to clusters leads to inferior generalization results. Topic modeling solves this problem and allows clear thematic separation by probabilistic assignment of documents to topics [37]. *Latent Dirichlet Allocation* (LDA) nowadays is a widely used topic modeling technique that was introduced by Blei, Ng and Jordan 2003 [39] including a detailed description of the algorithm.

Perplexity is a measure frequently being used in natural language processing for model evaluation. It is the determined by the model’s ability to generalize the underlying structure of a training corpus. The measure is particularly based on the logarithmic probability of a word w_d divided by the total number of words N_d in a document d within a test corpus M. Equation 2 shows the exact relationship.

The lower the obtained perplexity value, the better the generalization-capability of the tested model [39] [40].

$$\text{perplexity}(M) = \exp\left\{-\frac{\sum_{d=1}^{M} \log p(w_d)}{\sum_{d=1}^{M} N_d}\right\} \tag{2}$$

2.3 Geospatial information in social media platforms

This section describes the different aspects of social media as a source of geographically referenced information. It discusses the general characteristics of crowdsourced data and its applications for mapping. Also, the representativeness of social media as a basis for mobility-related findings is evaluated and typical interaction patterns are discussed regarding popularity and location-based functionalities.

2.3.1 Characteristics of volunteered geographic information

Traditionally, geographic maps have been generated by professionals using high-end tools. This typically involves governmental action and significant investments. In fact, the scope of data collection initiatives is often limited and aggregated data is typically not directly accessible for the public. Moreover, a fragmented data collection scope in terms of attributes, units and quality make it difficult to integrate locally sourced governmental data. Multiple open data initiatives, such as the platform *GovData* in Germany, aim at providing public access of authoritative information in easily exchangeable formats. With only about 6000 documents provided in the geographic platform section as of April 2017, no sufficient coverage is reached [41].

Simultaneously, crowdsourced *volunteered geographic information* (VGI) receives growing attention. As mapping is conducted by a large number of volunteers that contribute to an open database, the collection of data is very scalable and works with consistent data formats. However, the involvement of large numbers of volunteers makes VGI prune to special quality characteristics that are discussed in the following paragraphs:

Quality distribution Especially in highly populated areas, a better overall data accuracy is achieved [42]. It was found that the number of voluntary contributors increases disproportionately in urban areas [43].

Shifted quality assurance While traditional mapping procedures rely on quality control by experts who create the map, VGI data errors are generally only recognized by other contributors or the users of generated content [44].

Proximity focus The quality of mapped data depends on the local knowledge of the contributor, especially if the mapping process is highly manual. One example would be the biased creation of map data from aerial or satellite imagery if image errors are present or important sections are hidden by trees. Contributors tend to collect information for locations which are close to their usual habitation [43].

Limited training Geographic data being collected by volunteers is in general less reliable than mapping completed by contracted professionals [44]. Depending on the intensity of training and the mapping experience, the quality of results can be varied.

Representativeness Characteristics and perceptions of contributors do not necessarily represent the society as a whole. Certain information might be exaggerated or neglected depending on which aspects are perceived to be important by the contributors. Additionally, only a small share of the total number of VGI platform users contributes information on a regular basis [45]. Budhathoki 2010 [46] analyzed the activity distribution of an open mapping platform and found only 0.01% of the registered users to be very active while 70% did not contribute.

Potential malicious use Motives of users to contribute can vary and individual actions taken to worsen the quality of collected data can potentially take place. Due to the shift towards quality assurance by other contributors or final users, malicious data can be active in the database for a varied amount of time [42]. VGI platforms, for this reason, typically introduce contributor rankings that grant more autonomy for experienced users than for recently registered ones.

These quality characteristics need to be taken into account when using VGI for further analysis. Recent research addresses the development of semantic frameworks to evaluate the quality of user-generated map content [47] [48].

Social Media Geographic Information is a subgroup of VGI, which originates from social media platforms and is prune to special characteristics. In comparison to initiatives with the main purpose of collecting data, the information collected in social media mainly consists of byproducts related to communication-focused activity. Place entities are generated mainly based on location check-ins or in the course of users expressing their thoughts towards place-related topics on the platform. Therefore, the collected information is a direct representation of the users' interests [49]. This increases the severity of the previously introduced issue of information representativeness. With regards to the literature, there have been no in-depth social media studies of this aspect focusing on geographical data contained. Campagna 2016 [49] emphasizes that 'a novel analytics is to be formalized for the peculiar data models which make this type of information different from more traditional vector spatial datasets' (p.49). The authors recommend considering the spatial, temporal and a contextual dimension of the available data. Also, the value of multimedia contents is emphasized for analysis.

2.3.2 Mobility characteristics of social media users

In Germany, about 89% of all people have access and also regularly use the internet [50]. About 50% of all citizens use social media platforms [51] while

Facebook alone counts 37.9 million users in 2017 [52]. This reflects about 41% of all internet users in the country [53]. Social media sites show a steady growth over the last years of about 14% annual increase in user counts [50].

Figure 6 shows the population in Germany as of 2014 per age group and compares the user count of Facebook. For people between the age of 18 and 29, more than 91% of all individuals are active Facebook users. For the age group of 55 years and older, only about 17% of all people are represented on the platform. Data from different statistical travel surveys is combined to estimate the age-specific mobility demand particularly with regards to car usage. Based on the user penetration for Facebook, the total distance traveled daily by Facebook users as car drivers is calculated.

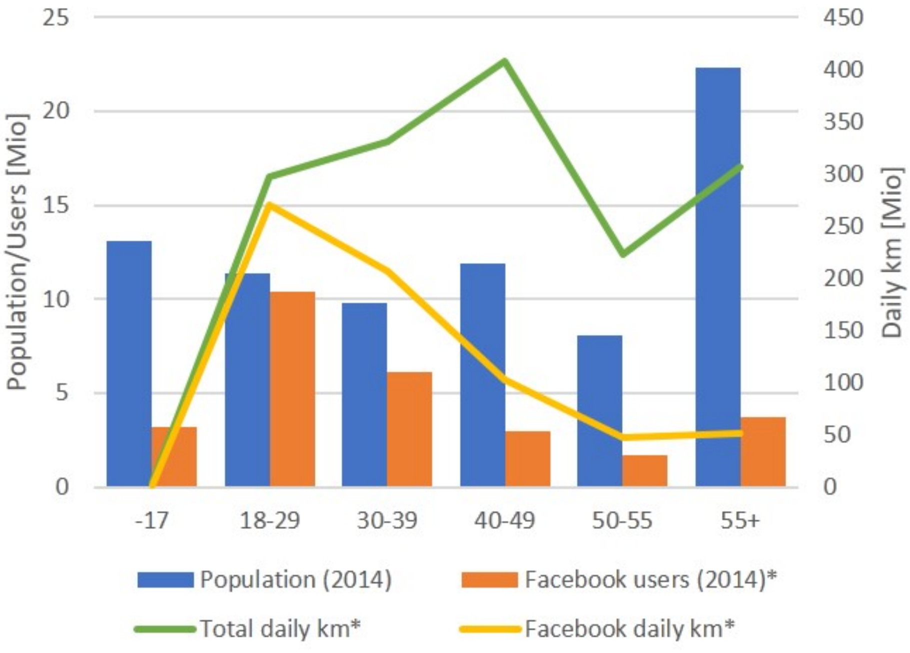


Figure 6. Mobility-specific representativeness of Facebook [54] [55] [56] [57]

The average daily travel distance per person in Germany amounted to 45 km in 2014 [56]. This mean is adjusted to age-specific values that range from about 31 km for people in the age of 17 years and younger to over 60 km for people between 30 and 39 years [57]. Subsequently, the share of distance traveled while driving a car [56] is multiplied with the age-group-specific population [54] and the daily distance traveled to obtain the total daily distance traveled by car. Using the Facebook user count instead of the population value leads to the estimate for the total distance traveled by Facebook users in Germany. Summing up, about 680 million km out of roughly 1,570 million km daily driving distance are caused by Facebook users which represents a share of about 43%.

The social media penetration is not homogeneous among all parts of Germany. The highest share is found in the state 'Saarland' with 88% of all people having social media accounts and the lowest share represents 'Thüringen' with 64% [58]. On a global scale, 56% of all Facebook users are male while the strongest gender difference is found in the age range of 18 and 34 [59]. Females are found to use Facebook more intensively while spending more time on the platform [60]. The respective data sources do not account for genders other than male and female. Furthermore, social media usage is observed to be independent from monetary income [61].

2.3.3 Interaction characteristics of social media users

To understand the offline mobility behavior of Facebook users, it is necessary to understand their motivation for online interaction with POIs and events. Unfortunately, this particular connection has not been sufficiently researched. A comprehensive literature review in 2012 by Wilson, Gosling and Graham [62] found that no research has been conducted regarding the behavioral drivers for user likes in general and particularly related to POIs and events.

With particular focus on humanitarian causes for likes, Brandtzaeg and Haugstveit 2014 [63] found this feature to be used mainly in the context of socially responsible liking. This concept describes the general willingness to support humanitarian organizations. Likes as an intermediate emotional reaction to the observed content were the second most frequent cause while the future access to further information was the third most frequent motivation. However, using the like feature is mainly seen a method for self-representation.

With regards to likes for company-representations on Facebook, access to information was found to be the main motivation for user likes. Moreover, access to special offers and other promotions were observed to drive user interaction. Showing support for the business to other users was also identified to be an important factor [64]. Further surveys confirmed these findings for brand representations [65].

Regarding location check-ins, Patil et al. 2012 [66] found users on Foursquare to share information mainly for self-presentation and access to certain social circles. Check-ins are regarded as a symbol of acknowledgment for a POI that helps support the user in being part of his or her social group. Thus, the motivation for online check-ins is explicable primarily with social and personal objectives while sharing the actual location is only a subordinate purpose [67]. Besides, it was highlighted that users may check-in to receive special offers limited to a certain geographic area [68].

2.4 Modal split in Germany

In Germany, about 3.4 billion kilometers were traveled per daily average for passenger transport in 2016 [69]. 80.1% of this were conducted by means of motorized private transportation. The rail sector was responsible for 7.8%, public road transport for 6.8% and air travel for 5.3% [70]. Considering data for the years 2005 to 2010, the car usage in Germany is about 4.6% higher than the European average of 75.6% (EU-27) [71]. 44% of all car owners use their vehicles on a daily basis while 32% use it at least several times per week. About 82% of all households have access to one or more cars while an average household has access to 1.4 cars [72].

The latter increases with the accessible income and only one percent of households with a monthly income of more than EUR 5,000 own no cars [57]. In average over all regions in Germany for 2016, 668 cars existed per 1,000 citizens [69].

Figure 7 provides an overview on the usage of transport means organized by different travel purposes. It is distinguished between mobility by foot, per bike, as car passenger, as car driver or via means of public transportation. The distribution of traffic generated by the presented modes is added as a line chart. All indicators are based on the assessment of traffic intensity as the total distance traveled while all numbers given represent shares on a percentage basis.

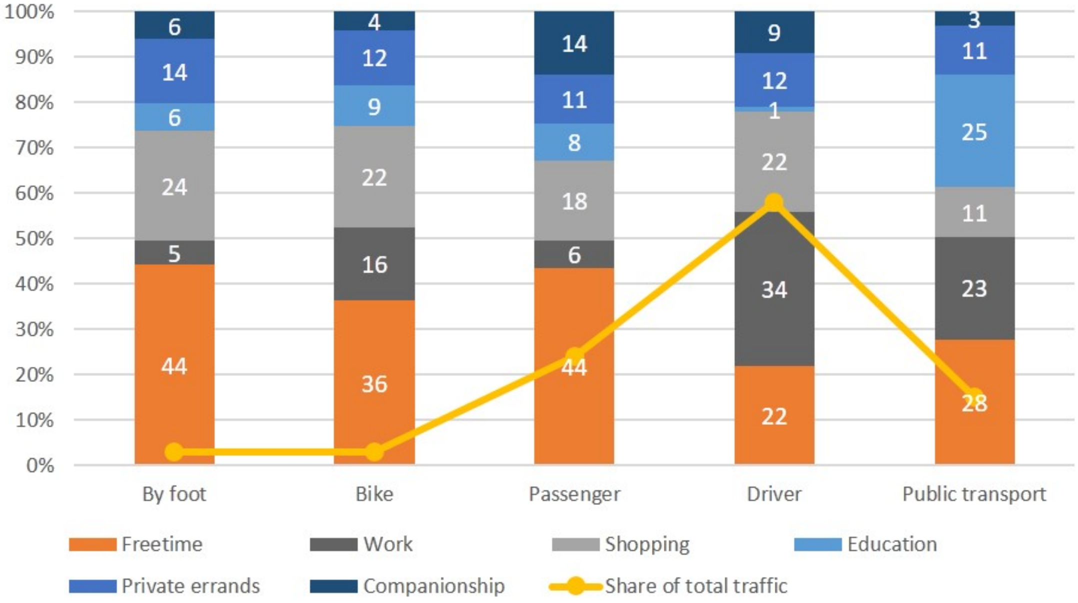


Figure 7. Modal split in Germany by travel purpose 2008 [57]

Leisure activities are responsible for the biggest share of traffic with regards to four out of five distinguished travel modes. Only for work-related trips, traveling by driving a vehicle is more popular. This travel mode makes up 58% of all traffic while traveling as a passenger is related to a share of 24% on the total distance traveled. Shopping trips are almost evenly popular with regards to all travel modes but public transport. The latter is especially popular for education-related travel.

The category 'private errands' sums up all activities that do not fit in any other group.

Among different age groups, there are also meaningful differences with regards to the means of travel chosen by individuals. Figure 8 distinguishes between eight age groups and the previously presented travel mode schema. While the distance traveled by other means of traffic, especially by foot, decreases until the age of 40-49 years, the importance of car travel rapidly increases. About 59 percent of all traffic created by individuals in this age group is caused by motorized vehicles. This correlates with the share of individuals who have car access being presented as a line chart. The required age for obtaining a driver's license in Germany is 18 years while accompanied driving can be done at the age of 17. This is why a very small share of distance traveled in driver role is conducted for this age group. Elderly people tend to reduce active driving and do most trips walking. This shift is supported by a decrease of the average trip length by about 50% in between the age groups 40 to 59 years and over 75 years [57].

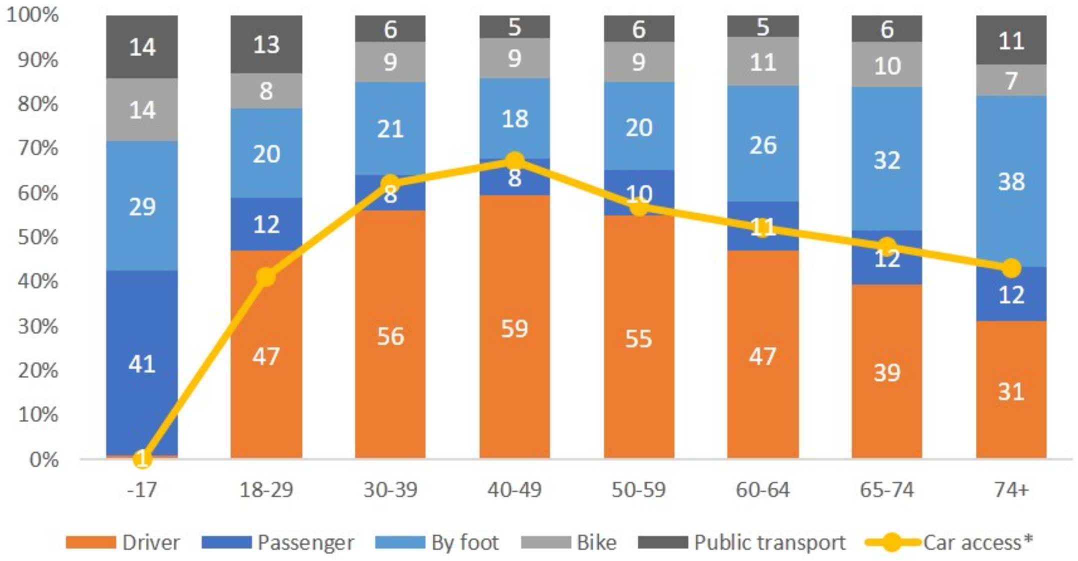


Figure 8. Modal split in Germany by age group 2008 [57] (* interpolated [56])

CHAPTER 3

Literature review

In this chapter, the state of the art in parking demand modeling is evaluated with focus on potentially relevant influence factors. Particularly, the impact of POIs and public events is analyzed in surrounding areas of parking spots. Also, a general summary of modeling activities with regards to travel mode choices and specifically, car usage.

3.1 Modal split modeling

In developed countries, people typically have the opportunity to choose from different travel mode options. The contributing factors that lead to a certain choice of mode alternative have been intensively studied in various countries. Mainly based on traveler surveys, researchers have distinguished macroscopic and microscopic influences. Macroscopic factors describe the overall setting where the mode choice is made and cover superordinate socio-economic and structural aspects. Microscopic influences relate to the traveler as an individual and the specifications of the planned travel. Factors such as age, income level or cost of mobility are covered as part of this subgroup [73] [74]. Figure 9 summarizes relevant influence factors in accordance with three new categories. It is distinguished between factors that are based on the characteristics of the decision maker and attributes of the specific trip made. Also, determinants related to the parking situation at the trip destination and factors considering travel mode alternatives, in particular public transport, have been emphasized in the literature. The provided collection of factors represents the intersecting set of findings derived from various studies in different countries. The annotated arrows indicate correlations between the respective factors and the stated preference of car usage as travel mode.

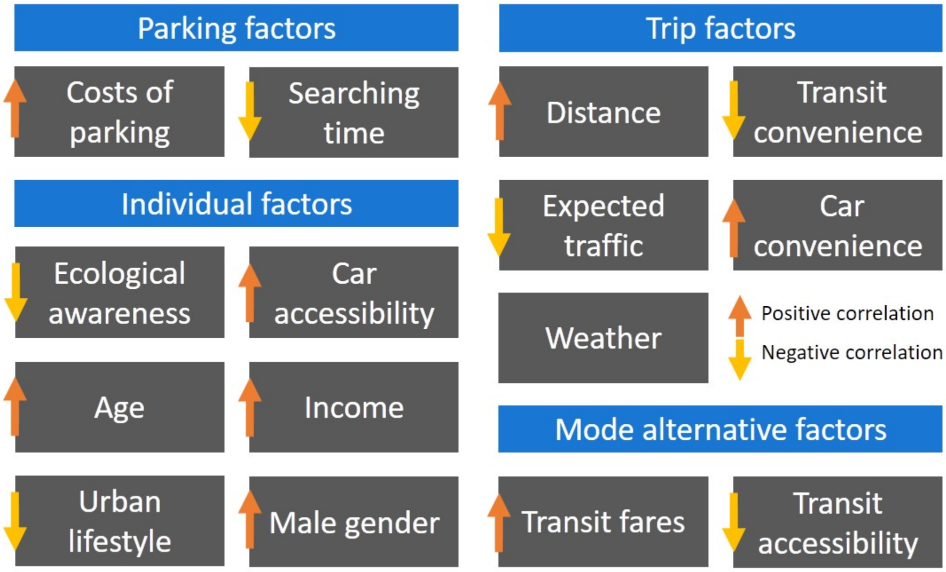


Figure 9. Relevant influence factors on the preference of car-related travel

Ding and Zhang 2016 [74] assumed that each individual chooses the preferred travel mode based on a maximization of perceived utility. As travel mode alternatives, respondents were asked to pick either public transport or car usage. Clustering algorithms were applied to form groups of respondents with similar individual characteristics. Gender, occupation, income and car ownership were considered as decisive factors. Ultimately, a set of generalized findings was derived from the obtained data using a multinomial logistic regression model. It was found that individuals with higher income tend to choose cars over public transport mainly for reasons of comfort. Respectively, travelers with comparably low income preferred to choose means of public transit. Moreover, even though long traveling times are perceived negatively, car travel was found to be generally more accepted.

Weis et al. 2011 [75] conducted a travel mode survey in Switzerland, indicating a significant effect of parking area characteristics on the travel mode. As a decision variable, the perceived utility of parking areas was used. It was found that males tend to use cars as the preferred mode of choice more frequently. Individuals with higher income showed less consideration of fuel prices and costs of parking. However, in general, costs of parking were found to be a highly relevant factor. Car accessibility had a positive effect on car usage in general while the necessary time for finding an available parking spot was perceived highly negative. In particular for short travels, the time spent searching and the monetary costs of parking were found to be important. Moreover, owners of public transport passes show a general preference for avoiding car usage while the particular travel is relevant for the decision making process. If many changes of public transport means are necessary or if significant waiting time is involved, this decreases the perceived mode utility. Finally, the authors point out that car usage is typically connected to time savings compared to alternative travel modes. The required duration of mode-specific travel, in fact, is another relevant decision factor.

Braun Kohlova 2016 [76] conducted travel studies for several cities in the Czech Republic. Focusing on travel mode utilities, tendencies towards car usage for males with high personal income and a superior work position were identified. Also, variables describing psychometric preferences towards certain lifestyles were found to be influential. The study included questions regarding the importance of travel accessibility and indicated preferences for short distances in a city lifestyle. Clark, Chatterjee and Melia 2016 [77] find ecological awareness to be an important influence factor that reduces the likelihood of car usage. Moreover, commuting distance, the residential context and the availability of vehicles within the household were identified to be relevant.

Badland et al. 2010 [78] examined work-related travel behavior in New Zealand and found trip convenience and travel mode accessibility to be crucial influences. Moreover, the availability and cost of free car parking, traffic intensity and the convenience of alternative travel modes, in particular public transport, were identified to have significant influence on the mode choice.

Furthermore, several studies have considered individuals' age when making travel mode choices. In most cases, it is implied that car usage is highly important for elderly in different countries [73] [79] [80] [77]. Böcker, Van Amen and Helbich 2016 [81] find elderly to rely on other modes of transport for the use case of the Rotterdam city area. The study compares a reference group to the behavior of the elderly. It is found that the spatial context of the planned travel also has an important effect. If the trip is set within a very dense urban environment, cars are less likelier to be chosen. The preference for car travel also positively correlates with the household size and the number of vehicles available. Moreover, it was found that individuals who showed certain preferences in the past are likely to continue making similar choices in the future [80].

Adverse weather conditions also have an empirically-proven effect on traffic patterns. Drivers tend to cancel or postpone trips, change routes, run errands preferably nearby and choose public transport instead of car travel if road congestion is present [82]. Rainfall is found to decrease car traveling speeds and the overall traffic volume. The corresponding effects for snow are similar but intensified [83] [84]. Winds and low temperatures are found to increase car and public transport travel [85]. Besides, Cools, Moons and Wets 2010 [86] highlight that the influence of certain factor combinations varies among different locations. This is explained by varying travel motives, for example those related to roads used mainly for leisure or alternatively commuting. While leisure is a rather flexible travel context where weather is an important influence, work-related travel is rather inflexible [85]. With specific regards to parking, low temperatures and rainfall were identified to increase the likelihood of drivers to choose off-street parking instead of on-street alternatives. The presence of snow, on the other hand, did not have a significant influence [87].

Yue, Cheng, Tai 2009 modeled the choice of travel modes related to urban events in accordance with a combination of socio-demographic and travel-specific attributes. The event popularity and resulting traffic demand were assumed to be known based on historical experience of the organizers [88]. This assumption is reasonable for recurrent events in established settings where a sufficient amount of data is available. The popularity of one-time events and non-commercial meetings is much harder to estimate. In fact, the study results are limited to events with required registration or ticket purchase.

Papacharalampous et al. 2015 [89] studied the modal split for big events in the city of Amsterdam. Pedestrian flows during an exemplary set of events were analyzed to calculate the share of visitors traveling with different travel modes. It was found that 74% of the visitors arrived by car with a person-per-car ratio of about 3.5. This extraordinary importance of automobile travel is interpreted as a result of the events' attractiveness to long-distance travelers and the convenient accessibility of the venue by car. Attendees have certain expectations regarding the expected parking search time at the destination and this factor does have an impact on the observed parking demand. As another factor, long-distance public transport in the case study offered comparably high group fares and limited traveling options after the considered events finished. This is interpreted as a secondary reason for the high car usage ratio.

No research focusing on smaller events, especially populated with social media data, has been conducted. Moreover, no findings are available that try to quantify the relevancy of interesting place characteristics for travel mode choices. As described above, the most present geographic level of detail being applied in current studies only aims at a regional perception.

3.2 Modeling of driver behavior

Collura, Fisher and Holton 1998 [90] investigated the behavior of drivers when searching for parking using simulated driving scenarios. Common search strategies of participants were naturally focused on decreasing the total travel time. Other studies highlighted that drivers primarily tend to circle in increasing distance to the desired destination [91], choose off-street alternatives or illegal parking spots [7]. Generally, on-street spots are preferred by drivers due to a more convenient access [14].

Certain *individual factors* have been identified that influence drivers' behavior and the observed parking occupancy patterns. These include the individual price sensitivity with regards to the available parking choices [92], drivers' knowledge of the area [91], required walking time and distance to the actual destination [93] [94]. Personal preferences of certain parking spots can also play a role in decision making [95]. Price sensitivity was found to vary among different city areas [96] being closely interconnected with the driver's travel purpose [97] [91].

Moreover, the importance of *contextual factors* is highlighted. This includes the population density in the destination area [96], the total number of parking spots available nearby, their parking turnover rate and the type of destination [91]. So-called *non-habitual* influences, for example, special events or traffic incidents, also need to be taken into account [98]. One part of the research community sees a strong interconnection between traffic volumes and parking occupancy [99]. On the other hand, several studies indicate that there are only minor correlations between parking and the observed traffic flow intensities in one area [22] [100]. Moreover, illegally parked vehicles can have an important impact [7].

3.3 Event influence on the parking situation

3.3.1 Estimation of event popularity based on social media data

Large-scale crawling and analysis of social media data for scientific purposes dates back to the mid 2000's with the structural network benchmark of Mislove et al. 2007 [101] being one of the most popular publications in the field. Nowadays, there are several frameworks that aim at analyzing social media streams to detect geolocated events. Popular examples are the emergency recognition framework of Xu et al. 2015 [102] or the traffic event classifier of Candelieri and Archetti 2015 [103]. Additionally, several high-level event identifiers were developed that provide a generalized applicability [104] [105] [106]. All of these include functionalities to make large-scale data sources accessible, to prepare sets of important attributes and to conduct data analysis tasks such as clustering or classification.

In the literature, four different types of gatherings triggered by social media are distinguished: Typically recurrent, scheduled meetings using specific event functionalities, planned semi-scheduled gatherings, ad hoc meetings and bigger, rarer special events [107]. An online-survey among 55,000 participants revealed that 58% of users on Facebook agree that online interactions also drive event attendance in real-life [108]. At the same time, web-based communication was also found to trigger actions in the online domain rather than having an effect on the offline behavior of individuals [109].

Du et al. 2014 [110] distinguish between three groups of influence factors that determine the event attendance in the context of social media: Content preferences of the users, the spatiotemporal context of the event and certain social influences. The authors developed a methodology to give users recommendions for future events based on the respective similarity to past events that have been attended. Using text mining, the similarity of events on a content-basis is derived from names and descriptions. Besides, the temporal similarity is calculated based on weekdays and time of the day. Lastly, spatial similarity is calculated as the distance between historical and recommended event locations. Later research used similar data representations and features to determine event similarity [111]. Figure 10 provides an overview on influence factors on the individual decision whether or not to attend an event promoted in social media. An adapted categorization is applied that comprises contextual, individual, social and event-specific factors.

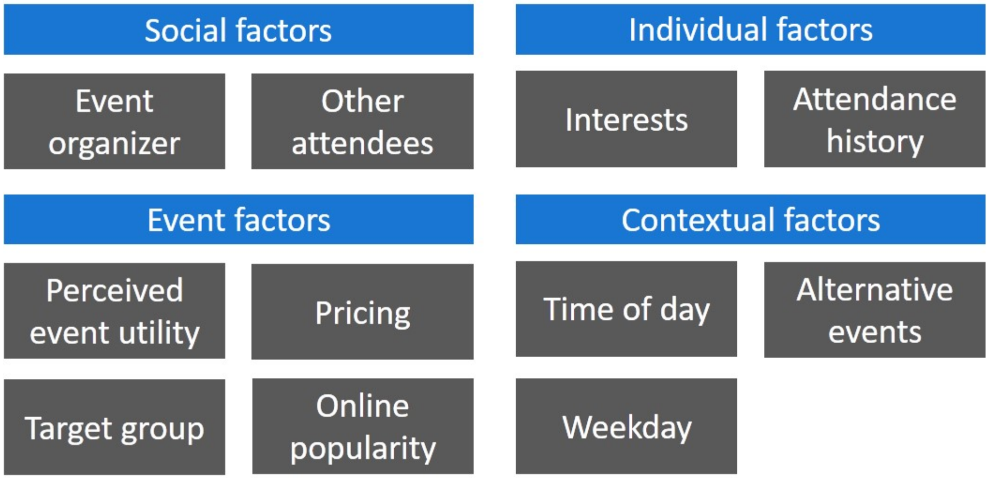


Figure 10. Relevant influence factors on event attendance

Bogaert, Ballings and Van den Poel 2016 [112] introduce a topology for event popularity modeling that is defined in accordance with the respective input data used. It distinguishes between published models that consider complex *network data* and approaches that focus on *user data*. Network data refers to interactions among users that serve as an indicator for social relationships and group behavior. User data comprises individual user information that reflects a specific platform usage behavior. The study highlights the benefits of incorporating network data for predicting the attendance of events based on a research population of about 950 users and about 2,500 events. An overall attendance rate of 78% was reported. The authors explain the relevancy of network data with a phenomenon called *endogenous group formation*, often also referred to as *homophily*. It denotes the preference of users to follow the decisions of their peer connections on social media. Research focusing on alternative social media platforms confirms the importance of network data for event popularity prediction [113].

Kawano, Yonezawa and Kawasaki 2014 [114] predicted event popularity based on the relationship among participants and their social media friends. It was found, that a relatively large number of users interconnected with event participants indicates a high popularity of analyzed event. A relatively low number of connections between external users and the attendees may be a sign of a limited target group and a lower attendance count in real-life.

Mynatt and Tullio 2001 [115] modeled the decision for or against one of many concurrent events using a Bayesian classifier on a person-specific level. The model included features such as the role of the individual in the event context, category and existence reminders to estimate the individual priority towards a specific appointment. Simultaneously, the user availability was estimated by locations of the event, information about the potential attendees and further contextual parameters. The model was deployed as part of a prototype calendar application and is intended to support collaborative work environments.

Paris, Lee and Seery 2010 [116] studied the technological acceptance of social media, in particular related to event information provided. The developed model considers the ultimate intention to attend an event as a consequence of several preliminary factors that must be fulfilled. Among others, trust with regards to the information provider, enjoyment and perceived usefulness were critical influences that affected respondents' choices. Slightly adapted models were developed by later studies [117] and the importance of trust and acceptance for user decision making was confirmed [112].

Michalco and Navrat 2012 [118] also identified critical factors that influence individuals' decision to attend events which are advertised in social media. Time of the event, the event organizer and other guests attending as a reflection of social bonds are considered for an estimation tool that predicts the likelihood of individual attendance. Later studies confirmed the importance of these factors in the decision-making process [119]. The authors also studied the relationship between claimed online event attendance and the actual real-life actions of individuals. The designed estimation tool performed with an accuracy of about 70% correct classifications. This supports the general applicability of likelihood-based methodologies in the area. The study does not provide estimations regarding a potentially generalized attendance rate for events in the social media domain.

As part of a report created for Facebook, Deloitte 2015 [120] spreads the assumption that about 50% of the positive or uncertain responses to events lead to offline attendance. Actual statistics related to more than 10,000 professional meetings held in Japan are provided by the event platform *Doorkeeper* [121]. It was found that mid-week events are more likely to be skipped as attendees are potentially busier with other activities than on the weekends. In the presented data, events on Saturdays have the highest check-in rates of approximately 85% whereas Wednesdays correspond to rates of only 78%. Also, paid events are more likely to have high attendance rates as individuals value the appointment also on a monetary basis. Moreover, smaller events were found to have a higher attendance rate compared to large events. The authors interpret this to be a reflection of the participants' obligation to share actual attendance information with the organizer. This effect is called *social loafing* and describes the belief that individual actions such as responding to event invites is not necessary for large invitee counts. In fact, related details in Social Media are a much less accurate representation of the actual event [119].

Huang, Wang and Yuan 2014 [119] examined the actual offline participation of individuals who indicated future attendance of certain events. Transcripts of interviews were used to extract common beliefs and behavior patterns. It was found that individuals perceive event invites on social media less binding and personal than direct notifications from the organizer via instant messaging. Furthermore, the ratio of accepted and declined event invites serves as an indication of the event attractiveness. Positively responding individuals tend to question their choice if there are more declines than accepts. Also, interviewees fear that their interactions with controversial public events affect the way their online profile is perceived by others. Moreover, certain non-literal meaning of the response options must be considered. Declining is perceived as an expression of contentwise opposition while accepting may serve as a symbol of group membership even though no actual attendance is planned.

3.3.2 Event influence on parking demand

Human mobility patterns are found to be highly recurrent and predictable when being analyzed on an aggregated scale [122]. The traffic created by special events, however, is more difficult to forecast. Most publications in the field focus on the estimation of event popularity and do not connect their findings to the implied traffic or parking demand. Typically, only extraordinarily big events receive special attention that may potentially lead to temporary traffic management by the responsible authorities. Here, past experiences are sufficiently available to serve as a basis for traffic planning. The popularity and mobility demand caused by smaller events, however, depends much more on contextual factors. These include aspects such as the social group addressed and the modal split present. Especially interactions of several simultaneous events in one destination area lead to complex traffic patterns that are hard to predict [123].

The first publication that highlighted the relationship between parking availability and local events was based on the manual characterization of standard occupancy patterns. Several influence factors for a chosen set of parking spots in the city of Munich were evaluated by examination of occupancy time-series over several months. The study found varying importance of certain factor combinations with regards to different parking spots. Their spatial setting was identified as an important distinguishing criterion. As interacting influences, weekdays, holidays, weather conditions and nearby events were examined. Similar occupancy patterns for parking areas with identical influences have been qualitatively stated. This indicates a certain predictability of on-street parking occupancy under consideration of the mentioned influences. However, no quantified examination of factor relevancies or methodological details of the applied forecasting methodology were reported in the available sources [124] [125].

Pereira, Bazzan and Ben-Akiva [123] modeled the arrival rates of public transport for urban events by extracting distinctive features from online data. The authors benchmarked several model alternatives and found ANNs to be the most accurate approach for modeling the changing arrival rates in the form of a time-series. The presence of deviant modal splits was identified based on the thematic character of the nearby event. This effect is interpreted to be a result of varying mindsets of the respective target groups. The study included information from social media, as well as several specific event websites and focused on event capacities of at least 1,000 attendees. The authors suggested the creation of more sophisticated features based on the available events data to improve the obtained model quality. The relationship between large-scale online event information and offline changes of the parking situation has not been studied.

3.4 Points of interest influence on the parking situation

Businesses, attractions and other places are often referred to as *points of interest* (POIs). They offer different interaction opportunities and attract people in various ways. This leads to urban traffic and ultimately also a certain demand for parking in the respective areas. In fact, the interaction schemes between people and points of interest, as well as their popularity, is hypothesized to be relevant for predicting the location- and time-specific occupancy of nearby parking areas.

Li et al. 2013 [126] analyze data from more than 2.4 million venues in different geographical regions from the social network *Foursquare*. The study characterizes the obtained dataset and finds superior popularity for places with highly detailed descriptions. In terms of category, especially food-related places show high popularities. The highest number of check-ins can be found with regards to places in the transport category. Repetitive visits occur typically in residential or professional places. Besides, older database entries are typically more popular.

Furtado, Fileto and Renso 2012 [127] use mobile device data to assess the attractiveness of places. The authors distinguish between several observed movement patterns to derive conclusions on the popularity. The number of stops at certain points of interest is used as a key popularity indicator.

Heterogeneous parking demand patterns have been identified for different areas of cities [128] [129] [99]. Most parking forecasting models presented in the literature account for this factor by implicitly considering a location-specific spatial context. These approaches distinguish between geographic cells and do not explicitly take into account places or land use [100]. POIs, in fact, are introduced only as an implicit side factor among multiple area characteristics.

Landry and Morin 2013 [130] found that the parking-related impact of interesting places follows a Gaussian distribution. Multiple aerial images of parking areas at certain times served as a source for the observation that drivers prefer to park as close as possible to the respective POI. The study focuses on the occupancy within large parking lots and does not consider urban environments on a larger scale.

To conclude, the relationship of interesting places and changes of the parking situation has not yet been analyzed in an isolated manner. No sufficient research exceeding POIs as part of larger geographic areas has been conducted with regards to the effects on urban parking demand.

3.5 Modeling parking occupancy

Parking availability modeling on a generalized level can be subdivided into statistical approaches and methodologies based on artificial intelligence [131] [132]. Here, another category is introduced that focuses on simulative approaches. For all groups, the most common application is short-term occupancy forecasting of *off-street* facilities such as parking garages or paid areas. As sufficient data is typically available for this group, time-series forecasting is widely applied with high accuracies.

Statistical models In terms of statistical models, popular approaches are ARIMA [133] and non-parametric [22] [99] or geographically weighted regression [96]. Caliskan et al. 2007 [134] developed a Markovian model and tested it against occupancy data from a traffic simulation. Ziat et al. 2016 [135] developed a combined time-series model for traffic and parking availability that was found to outperform other modeling approaches. This supports the intrinsic connection between both parameters. The system was based on the vectorial representation of roads and off-street parking areas in a common large-dimensional space. Chang 2016 [136] used linear regression to model the occupancy of paid on-street parking spots.

Simulative models Different simulations have been developed that derive overall occupancy information from the aggregated driver behavior. Agent-based modeling was successfully applied to test the impact of parking policies [137]. Caicedo, Blasquez and Miranda 2012 [138] applied a discrete choice model to estimate the allocation of parking requests, future departures and the expected availability of parking. The developed model depends on an initial set of calibration parameters that reflect the underlying driver characteristics in the modeled area.

Artificial intelligence As the most popular tool within this group ANNs with backpropagation architecture have been applied for off-street occupancy [139]. ANNs are especially useful for applications where large amounts of input data are available and little is known of the actual relationship among the input factors. Moreover, they show advantages for modeling non-linear relations and factorial interactions [140][100]. Chen 2004 [141] benchmarked several forecasting models in the parking domain and found ANNs to outperform regression-based approaches. Richter, Martino and Mattfeld 2014 [142] applied spatio-temporal clustering techniques and found significant potentials for saving computational resources while facing only minor forecasting quality decreases. Ji et al. 2014 [131] generated occupancy forecasts using wavelet neural networks and a concept based on the largest Lyapunov exponents. Vlahogianni et al. 2015 applied a combination of ANNs and survival models to predict the availability of monitored on-street spots [143].

Pflüger et al. 2016 [100] used a backpropagation ANN with sigmoid activation function to forecast the availability of on-street parking in the city of Munich, Germany. The authors focused on publicly available data as model input data. Validated occupancy information was generated by volunteers using a smartphone application that supports the manual tracking of available parking spots. Subsequently, *feature selection* was applied and time, location and weather were identified as the most crucial factors. Only slight increases of the prediction performance were noted due to factors related to traffic and local events. The generated models achieved a forecasting Mean Average Percentage Error (MAPE) of about 16%.

Modeling on-street parking occupancy still remains a widely unsolved problem. The published approaches are not able to derive generalized models with sufficiently high prediction quality. No ubiquitous application exists, that adequately explains the observed variance in real-life occupancy data.

CHAPTER 4

Data Acquisition

This chapter discusses the methodologies developed for collecting large-scale POI and event data from Facebook and a selection of reference data sources in the respective fields. A preliminary evaluation of these online platforms limits the scope of the investigation. Moreover, exploratory data analysis is conducted for the obtained datasets, revealing certain quality-related and thematic characteristics.

4.1 Preliminary evaluation of leading online data sources

Detailed evaluation of databases for subsequent processing and analysis requires full access to the provided information. This typically corresponds to significant efforts for the development of web-interfaces and crawler modules. Therefore, a preliminary evaluation of potentially relevant data sources is conducted based on publicly accessible platform descriptions and reported indicators. Table 1 provides an overview on leading community-based online data sources that offer large-scale information related to points of interests and events. The thematic focus of the respective platforms (social media, mapping and events) is distinguished.

	Social Media	**Mapping**	**Events**
	Facebook	**OpenStreetMap**	**Eventbrite**
Users [mio]	1860 (2016) [144]	2.8 (2016) [42]	20 (2012) [145]
Alexa ranking [146]	3	6828	768
Basic data	x	x	x
Popularity data	x		
License	Proprietary	ODbL	Proprietary

Table 1. Leading online data sources in relevant areas

As of 2017, the leading online social network is *Facebook* with 1.8 billion monthly active users worldwide. This represents 22% of the global population.

In North America, more than 68% of all people use the platform [147]. In 2016, the global user count increased 16.9% while the company generated revenues of USD 27.6 billion [144] and controlled about 16.2% of the global online advertising market [148]. Besides major products focusing on communication, Facebook has increased its commitment to geographically-driven functionalities. This includes extensive user-generated databases of POIs and public events, that can be accessed by third-party app providers using Facebook's *Graph API* [149]. Using the public API, static information, such as name, location, or category for both, places and events, can be retrieved after obtaining an access token via complementary registration. Also, specific social media features are provided that indicate online popularity. This includes counts of fans, check-ins and user ratings. Facebook's database concerning points of interest is not fully based on voluntary contributions. Enhancement of the provided information is achieved by commercial licensing of third-party content [150].

Foursquare, as an alternative social network offering location-based services, only had about 55 million active users in 2015 [151]. In the *Alexa* internet ranking, a popular information service focusing on website popularity, Foursquare only ranked near position 1,900 while Facebook was among the top three most visited and interacted sites as of April 2017 [146]. Therefore, Facebook is chosen as the primary data source for geographic information extracted from social media.

Other than social media, extensive open platforms exist that focus on creating freely accessible maps with global coverage. Among other information, these projects typically contain a broad collection of static information related to POIs. One major source is *OpenStreetMap* (OSM) with 3.6 million users [152], worldwide coverage and about two billion contributions already in 2010 [153]. OSM contains millions of tag-based entries for points of interest all over the world in various categories. Users actively contribute to the collection of geographic information as this is the platform's main purpose. The actual OSM user count, in fact, is not as important due to higher contribution activity of individual users. Moreover, the comparably low Alexa rank (Table 1) does not genuinely reflect the platform popularity. As the accumulated map data is republished daily to be self-hosted by third-party application developers, not all web traffic accessing OSM data is recorded in this ranking.

An alternative mapping platform is *Wikimapia* with about 1.9 million users in 2013 [42] and a total of twelve million contributions in 2010 [153]. Contrary to OSM, geographic entities are partly cross-referenced with further information found online. According to the Alexa website ranking, the user activity on Wikimapia decreased in recent years significantly. As of April 2017, the website ranks near position 4,500, having lost about 1,500 positions since the previous year [146]. As the collected data is not as openly shared for third-party hosting as for OSM, the Alexa ranking accurately represents the actual platform popularity. Therefore, the overall extent and quality of OSM data is expected to be higher and it is chosen as the primary data source for POIs.

With regards to public events, diverse platforms exist that commercialize the aspects of marketing and ticket sales. In this field, *Eventbrite* is one of the leading suppliers. The platform had about 20 million active users in 2012 [145], representing the most recently published user-related data. According to the Alexa traffic statistics, Eventbrite ranks among the top 1,000 websites (Table 1) while reporting gross ticket sales of USD 2.0 billion in 2013 [154]. In contrast to event entities in social media, Eventbrite focuses on events that require users to give attendance feedback (RSVP), particularly paid events. Facebook also uses an RSVP system for managing the event database but feedback is less binding and no sales are conducted directly over the platform [119].

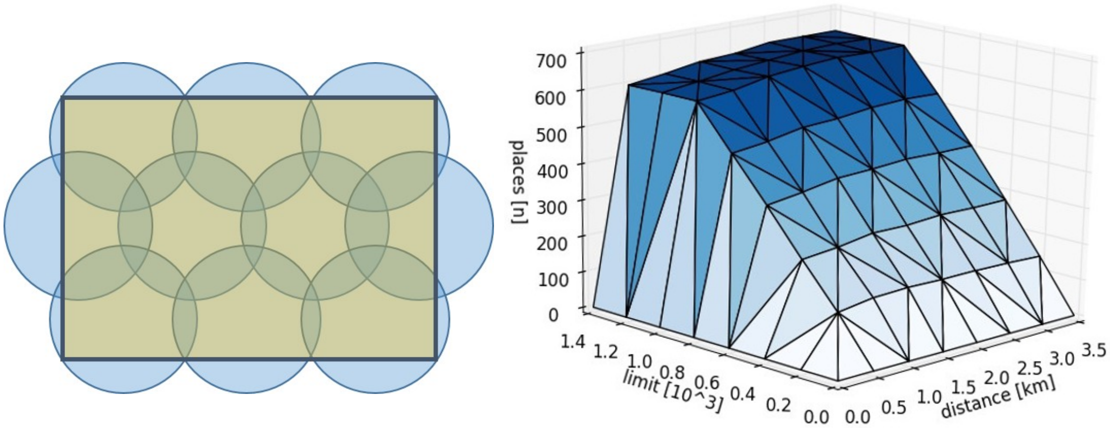
One relevant competitor with wide-spread coverage is *Ticketmaster*, offering tickets for about 230,000 events in 83 countries as of 2017 [155]. Ticketmaster generated USD 7.2 billion in revenues for 2015 [156] and maintains a stronger focus on large-scale events than the previously introduced platforms. As Eventbrite and Ticketmaster cover different segments of the event platform market, both must be taken into account for further analysis.

4.2 Facebook POI data

4.2.1 Data acquisition

Facebook offers place-related data via location-specific calls of its Graph API (v.2.8). Using a valid access token, places in the surrounding of a given geographic location are provided as a web service response. The information is received in the Java Script Object Notation (JSON) format. The API accepts longitude and latitude of the requested location, a 'distance' parameter specifying the size of the covered area and a 'limit' attribute that defines the maximum amount of objects to be responded. Larger requests take longer to process, use a higher data volume and create significant utilization of the data source. As the API is mainly used within third-party mobile applications, the data volume used and the request response time are critical factors that affect the perceived service quality. Thus, the limit parameter is used to reduce the amount of information requested. If large web service responses are not required for good user experience within the target application, mobile data consumption and API can be reduced in this manner. In the context of this study, the API is used for collecting nationwide geographic information. A parser module is necessary to retrieve the available place information in a systematic manner. Figure 11(a) provides a schematic overview how request locations are distributed to cover a coherent rectangular area. The position of retrieved place objects follows a radial pattern defined by the distance parameter chosen.

As a preparatory step, the API characteristics are examined by testing its capabilities for the city area of Berlin. Figure 11(b) shows the response characteristics for requests on the city center coordinates. The previously described request parameters ’limit’ and ’distance’, as well as the number of retrieved place objects are taken into account as plot axes. It can be seen that a implicit response limitation is reached between 600 and 700 place objects independently from the passed request parameters. The limit is reached already at a chosen distance of 0.5 km because the density of place objects for the test area is extremely high. In fact, greater distances passed do not lead to more POIs received. It has to be noted that about one percent of the objects received are located further from the request location than defined by the distance parameter. For further analysis, this phenomenon is neglected and a radial coverage is assumed.

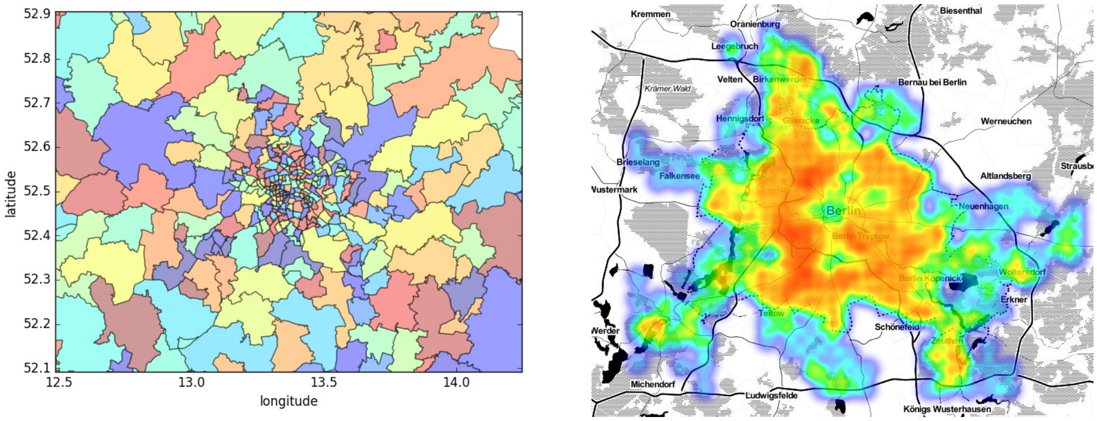


(a) Covering rectangle with radial API calls

(b) API performance for the city of Berlin

Figure 11. Facebook API request characteristics

In order to avoid loss of information, the request area size cannot remain static. The API call locations passed must be dynamically adjusted in accordance with the degree of urbanization for the target area. Areas with higher POI density must be covered with smaller areas and a larger number of requests. As an indicator for the density of POIs, zip code area sizes are considered. It is observed that the smaller the zip code area, the higher the degree of urbanization. Figure 12(a) shows zip code areas as colored polygons. Highly populated zones are indicated by a large number of small polygons while suburban or rural areas are covered with large zip code areas.



(a) Zip code areas as colored polygons

(b) Heatmap of API requests

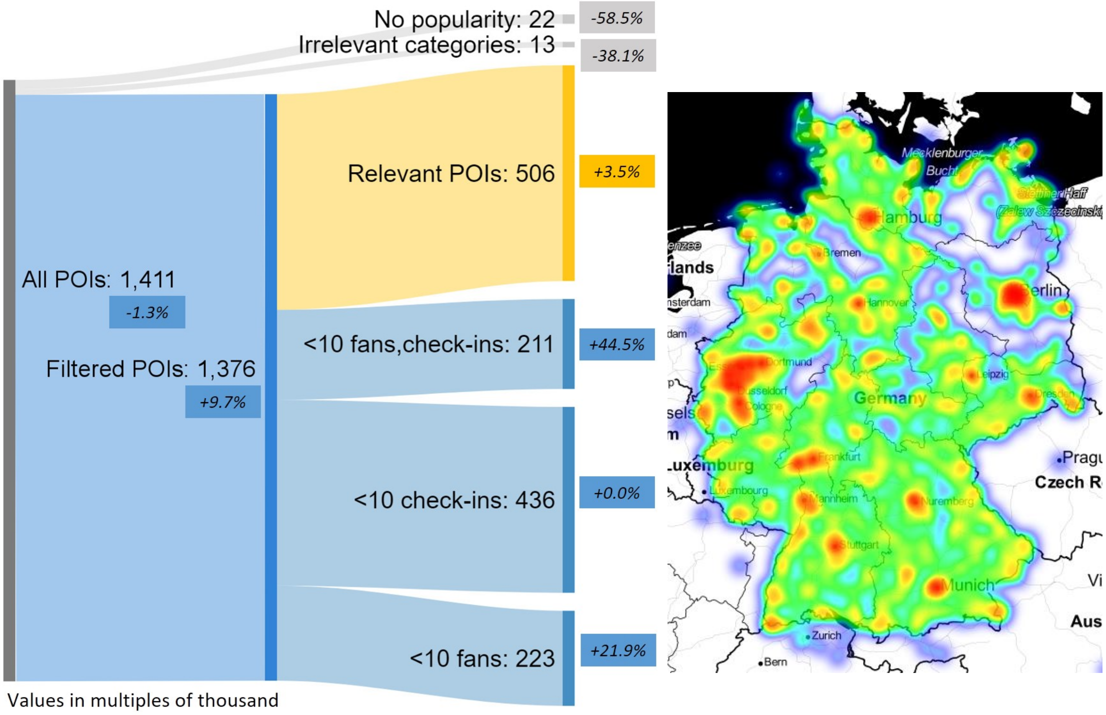
Figure 12. API requests per zip code area for the city of Berlin

API performance evaluations are conducted for further arbitrary locations in Germany to define the optimal relationship between zip code area size and API request distance chosen. For example, the contextual structure and degree of urbanization for the city of Braunschweig indicates an optimal distance between request points of about 2.5 km. Further tests in suburban settings allow request distances of 4.5 to 6.0 km before the API response limit is met. All zip code areas in Germany are clustered by size and are matched with request distances that are found to be optimal for the tested areas. This results in groups of equal zip code area size being covered by homogeneously spaced request points. In total, ten different clusters are distinguished. Figure 12(b) highlights the distribution of request locations for the Berlin region using a heatmap. The heterogeneous structure is obtained due to the large number of zip code areas involved. Especially in the dense center section, a high number of API requests is indicated by red coloring. Appendix F1 shows a histogram of the size distribution among all zip code areas. A boxplot highlights the fact that all areas lie within the 1.5-fold interquartile range (IQR) while the proportion of small, urban zip code areas is high. This can be explained with the generally high degree of urbanization in Germany. In 2015, more than 75% of the population lived in cities [157].

Establishing nationwide coverage in the presented manner requires about 47,000 API requests. Due to load limitations on the data provider side and high resource consumption for data processing, the number of requests must be minimized. As the ultimate purpose of AIPARK is the improvement of the parking situation in cities, service coverage is required specifically for dense urban areas. Thus, API calls for rural regions do not necessarily need to be conducted. 20% of all zip code areas represent about 40% of the total aerial extent of Germany. These are neglected in accordance with the pareto principle to reduce the number of required API call proportionally. Using the dynamic determination of geographic distances in-between API call locations, a total of 1.41 million place entries for all of Germany is obtained. Using a static distance of 2.0 km for the request distribution, only information regarding 1.38 million places can be collected. This represents a 2.2% surplus of aggregated data.

4.2.2 Data characteristics

Before the collected POI objects can be used for further analysis, an exploration of the obtained data and certain preprocessing steps are necessary. An overview on the distribution of objects in accordance with their respective popularity is presented in Figure 13(a). The annotation boxes contain relative values that indicate changes in the respective popularity segments within a time frame of 20 days. For generating this data, two separate datasets are obtained and object counts in the respective categories are compared to evaluate the dynamics of Facebook as a data source for geospatial information. The presented absolute popularity values refer to a state of the dataset retrieved in April 2017. The data source is found highly dynamic with regards to fluctuations of the number of POI objects retrieved for each category. Within the considered time frame, the overall number of POI objects slightly decreased while the increases are observed regarding popularity.



(a) Popularity distribution

(b) Geographic distribution

Figure 13. Facebook POI characteristics

About 22,000 out of 1.41 million POIs are considered to be not popular. No online user attention is centered on these objects with regards to the interaction measures likes and check-ins. It is assumed that this also corresponds to no real-world actions, especially in terms of trip planning and intended visits. Thus, these POI objects are neglected for further analysis.

In total, the dataset comprises POIs connected to 1,124 different categories. By manual identification, eleven of these are found to be structurally irrelevant for forecasting parking occupancy. These refer to rather large areas other than specific locations and intermediate effects on the parking situation cannot be reliably derived from these objects. Example categories are 'city', 'continent' or 'ocean'. The full list is presented in Appendix 8.6. Even though these POIs have a broad geographic relevancy, they are referenced to specific locations, leading to a bias in popularity for this area. For instance, objects in the category 'country' show a median fan count of more than 79,000, representing the highest popularity among all POI categories.

About 870,000 POI objects in the dataset only show a relatively low popularity. While 16% of the pre-filtered entries have less than ten fans, 32% have less than ten check-ins. About 211,000 POI objects have both low fan and check-in counts. Thus, only about 36% of all collected POI objects are particularly relevant for further analysis. However, for dense metropolitan areas, the interactions of low-popularity POIs are still assumed to have a relevant, combined effect. For this reason, these objects are also taken into account for further analysis.

Figure 13(b) shows the geographic distribution of filtered POIs in the form of a heatmap. Dense metropolitan areas, such as the cities Berlin, Munich or Hamburg, can be reliably recognized by red coloring. Excluded zip code areas with regards to data acquisition account for empty map areas. Among all POI objects retrieved, only about 25% contain information related to opening hours.

Figure 14 shows the ten most popular POI categories in the retrieved dataset. The number of fans for the respective places is presented as category-specific boxplots on a logarithmic scale. The length of the boxplot whiskers is set to one IQR. Additionally, the number of place objects per category, the median check-ins per object and the median number of users talking about the POI are displayed. No quantitative object count outliers are identified while the categories 'Government Organization' and 'Movie Theater' are most widely present. All categories show a strong variance regarding user interaction. The median value's robustness against outliers is more beneficial for ranking compared to the simple mean value. A limited number of POIs with extremely high like count are found for all of the considered categories.

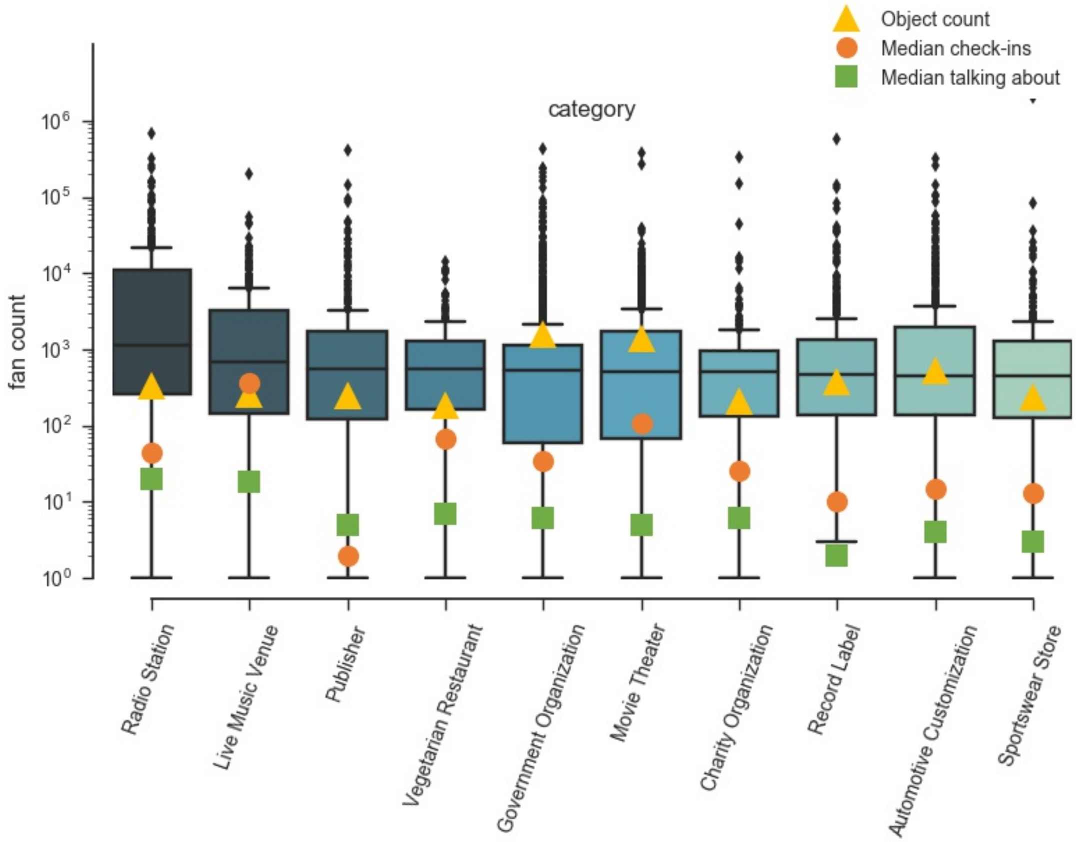


Figure 14. Most popular Facebook POI categories

The online popularity factors 'check-ins' and 'talking about count' are lower than the number of fans in all of the presented cases. Especially for the POI category *Publisher*, the discrepancy between fan count and the other popularity measures is very high. It is assumed that this phenomenon is caused by the fact that publishers are media organizations that are relevant not only in a certain location, but on a geographically broader scale. This effect can be stated for all POIs that interact with society independently from a specific physical location. For example, this is the case for POIs that represent strong brands or office locations of popular companies. Thus, the differentiation between *intellectual popularity* and *physical popularity* is necessary in social media. While intellectual popularity is fully based on the interests of users, it is independent from actual physical presence of the user at the POI. Intellectual popularity attributes for the example of Facebook POIs are 'fan count', 'talking about count' and ratings. Physical popularity, on the other hand, is directly connected to the presence of users at the POI and involves a mobility pattern related to it. By definition, this category has more inherent relevancy for the traffic and parking demand created by the POI. However, it has to be noted that online check-ins only occur if users actively decide to share their location. As the underlying motivations can be varied, physical presence is represented only to a certain degree and are directly influenced by the interests and attitude of users.

4.3 Facebook event data

4.3.1 Data acquisition

Over its Graph API (v.2.8), Facebook provides public access to open information created by users of event-related functionalities. Users can freely create private or public event objects and invite other users. This provides a platform for community-based interaction related to events. Commercial organizations mainly use these functionalities for marketing purposes and private users for facilitation of the event organization process. The Graph API provides two options for collecting event information in the form of JSON responses. One method is based on the previously obtained POI objects. Within Facebook's graph structure, these are directly linked to event entities as most of them are conducted at specific physical locations in contrast to online events. The second data acquisition methodology is a keyword-based search approach.

For the first collection method, the total population of place identification numbers is divided into batches of 50 pieces that represent the maximum accepted by the API within one request. These keys are passed with the API call to obtain events that correspond to the respective places. One separate parameter determines the time frame covered by the request. Facebook also allows retrieval of information related to past events. For a certain focus area, if more data is available that can be passed in one response due to internal performance reasons of the API, the retrieved file contains links to further response pages. These are exhaustively called with the crawler script to obtain all information that fits the specified parameters. As identification numbers for places frequently change, the script also provides functionalities to obtain updated object keys passes with error responses. This avoids mistakes in the crawling process. Data collection for all available places requires about 65,000 API calls while 57% are related to further response pages. In total, about 1.7 million event objects are collected by conducting this procedure.

The keyword-based approach uses a different endpoint of the API that is not location-specific. Thus, primarily keywords are passed that refer to certain locations. First, a list of 109 cities in Germany is extracted form a publication of the *Organisation for Economic Co-operation and Development* (OECD) [158]. These are used as keywords for the event search and lead to 14,700 objects retrieved in total. Mainly due to duplicate city names on a global level, only about 90% of these are located within Germany. Due to the comparably small number of objects retrieved, this approach is neglected. As a second procedure, specific POI names are used as keywords to collect events objects that have no formal connection in the graph structure but potentially show a real-world relationship based on syntactic similarities found. Testing a sample of 2,000 random place names, in total 5,600 event objects are returned while only 1,100 of these have sufficient location information and are set within Germany. It has to be noted that the keyword-based search requires user-specific authentication and the response volume is strictly limited. Thus, for large-scale data acquisition an a nationwide scope, this approach cannot be realized.

4.3.2 Data characteristics

Users on Facebook have different options to indicate whether they will attend an event. They can specifically accept the invitation of another user by claiming attendance. They can also note that they are just interested in the event and unsure of the actual attendance. Finally, they can also formally decline or avoid replying to the event invitation. User counts for all of these options are provided as attributes for each event in the dataset. However, due to policy changes as of April 2017, the count of users who declined a specific event is no longer available using the Graph API. By default, response objects with this attribute equal to zero are returned. Thus, this factor is neglected for further analysis.

Figure 15 provides an overview on the distribution of popularity indices and data quality. About 517,000 events from the total count of about 1.7 million objects are significantly popular. The number of attending and interested users is equal to zero for this segment. Thus, about 30% of the entire dataset is irrelevant for further analysis. A lack of online popularity is assumed to be reflected in real-life attendance and traffic demand created by the event. Objects without geographic coordinates cannot be referenced and also must be excluded from the dataset.

A smaller number of objects falls under both categories, lacking popularity and specific location information. Together, both represent about six percent of all events. For the creation of an event object, a specified starting time is required. Regarding event end times, a meaningful share of users tend to avoid specifying them. Roughly 271,000 objects do not contain end time information. For these, it is not possible to draw conclusion on the observed traffic patterns as it remains unclear which effects can be assigned to the event. Only about 47% of the collected data fulfills all formal data quality requirements while about 484,000 objects in this segment show particularly low popularity indicators. For public events, a number of less than ten positive attendance or interest responses are considered to be particularly low. Thus, only about 19% of all events are perceived to have immediate relevance for further analysis.

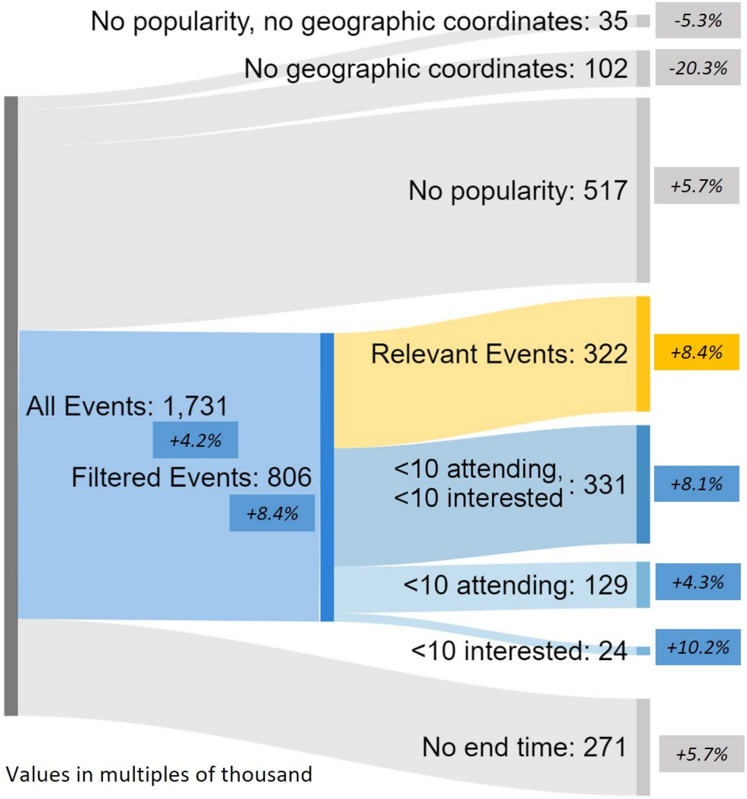


Figure 15. Facebook event data quality and popularity

The dataset comprises 40 different event categories. Only about 91,200 events have assigned information in the respective field, representing a share of about five percent compared to all objects collected. A strong tendency of the dataset towards events in the entertainment sector is observed. Work-related events are strongly underrepresented and only make up about 1.5% of all objects in the dataset. Figure 16 visualizes the event distribution based on the number of retrieved objects. Here, the 40 basic event categories in the dataset are assigned to eleven superior segments represented by homogeneously colored tiles. Subordinate tiles represent occurrences of the original categories. The segment for miscellaneous events comprises categories that may include different thematic trends which cannot be clearly assigned to other segments without introducing a bias. The explicit assignment of original to superior categories is visualized in Appendix F2.

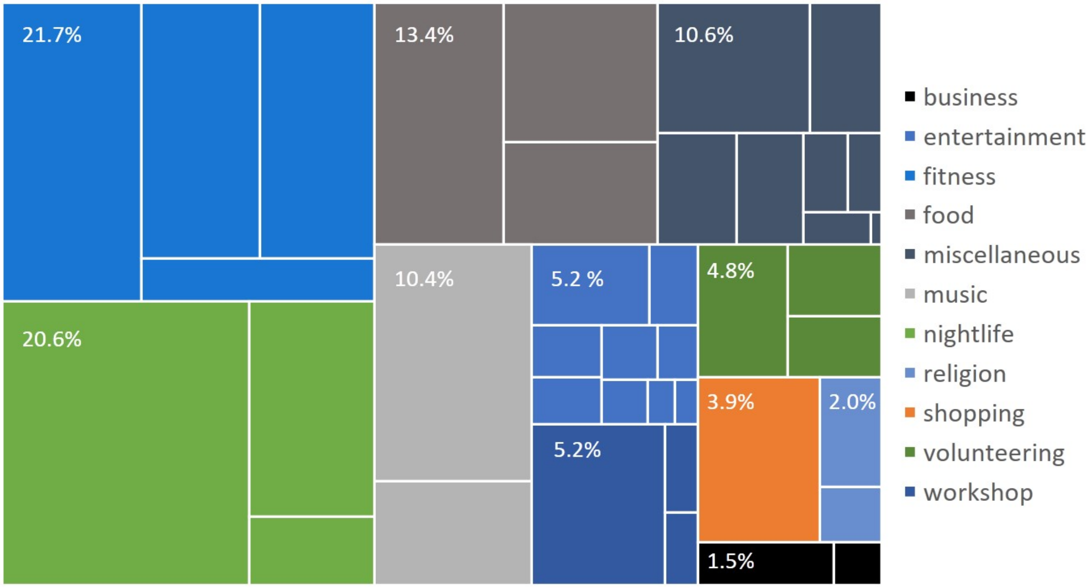


Figure 16. Distribution of event occurrence by superior categories

Figure 17 shows the ten most popular subordinate event categories sorted by the median attending user count. Nightlife-related events represent the most popular segment. The median interaction counts for interested users and participants who have not replied to the event invitation are located in an intermediate range. Nightlife events alone correspond to 24% of all attending users summarized over all categories. All segments contain single objects that represent characteristic attending count outliers. They are indicated by plotted points above the boxplot whiskers.

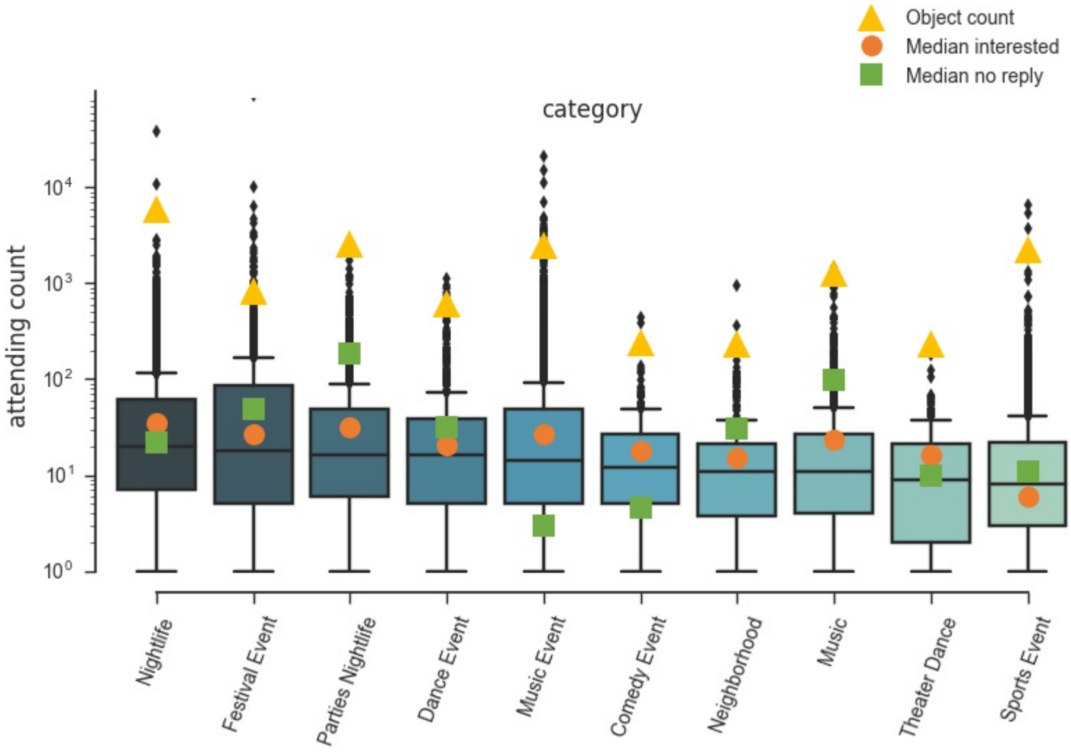


Figure 17. Ten most popular Facebook event categories

7.9% of the collected event objects take place in the future. This is expressed in Figure 18 by showing the accumulated share of event objects over a period of several months relative to the data collection time. In average over all event categories, the accumulation of past events approximates a linear function. This spans a time period of about 14 months prior to the request time. For future events, the accumulation reflects a near-asymptotic behavior. Only about two percent of the events are planned more than three months ahead of time. It has to be noted that the time distribution varies among different event categories. In

Figure 18, nightlife and comedy events serve as an example for differences in the timewise distribution. Generally, while nightlife events are rather planned on short notice, comedy events are prepared longer in advance.

Comparing the results of different data collections, about 5,400 events objects show changed popularity values after they had already finished. User interactions are technically still possible for this timeframe and are conducted to a limited extent. This stands in opposition to the initial assumption that past events are no matter of interest. In fact, when past event information is collected at a certain point in time, the obtained popularity indicators may slightly differ from the original values at the time of the event. However, as less than one percent of the filtered event objects is affected, this phenomenon is neglected for further analysis.

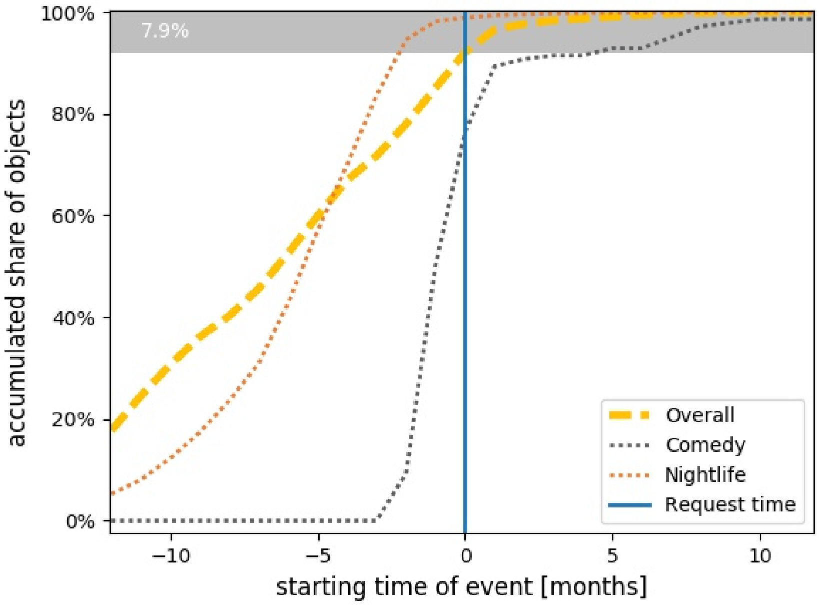


Figure 18. Facebook event distribution relative to the request time

4.4 Reference data

4.4.1 OpenStreetMap

The entire OSM dataset is continuously updated and provided as an XML file that offers global mapping coverage. Also, country- or continent-specific sections of the dataset are offered as smaller files that are easier to process from a computational perspective. Real-word entities are mapped using *nodes* that contain the object location and an identification number. Streets and paths are mapped using a list of nodes described by the *ways* data type. All map objects are specified using *tags* to capture the object characteristics. Each tag is included using specific pairs of keys and descriptive values. For example, a street within a residential area is represented by the tag 'highway=residential'. Tag usage is regulated by a set of community standards. In total, about 750 different key-value pairs are distinguished [159].

OSM serves a general mapping purpose. The data includes a broad variety of tags that do not only focus on POIs but can refer to all potential objects in the physical world. 143 tags are manually chosen that refer to entities that are expected to have an influence on traffic patterns and the parking situation. Tags denoting larger areas or nodes with a broader geographic relevancy, such as cities, are removed. The OSM dataset for Germany contains about 1.8 million objects that contain relevant tags.

4.4.2 Eventbrite

Eventbrite offers information related to both paid and free events. If a geographic location is passed to the API (v.3.0), it supplies objects that are located in the intermediate surrounding defined by a separate extent parameter. The output is limited to events that are currently marketed on the platform. Past event data cannot be retrieved. The data collection methodology is applied for obtaining

Facebook POI objects, as described in 4.2.1. However, for the case of Eventbrite, a fixed request radius is chosen as the API's paging functionalities prove to be fully stable in preliminary tests. This means that identical responses are obtained when a rectangular test area is covered either with multiple radial API calls or with only one request covering all of the considered area. As one response can only obtain 50 event objects, a corresponding number of calls to different response pages is needed. Request locations are generated with a fixed geographic distance to cover all of Germany. For the time of the conducted request, about 8,300 event objects in total are collected.

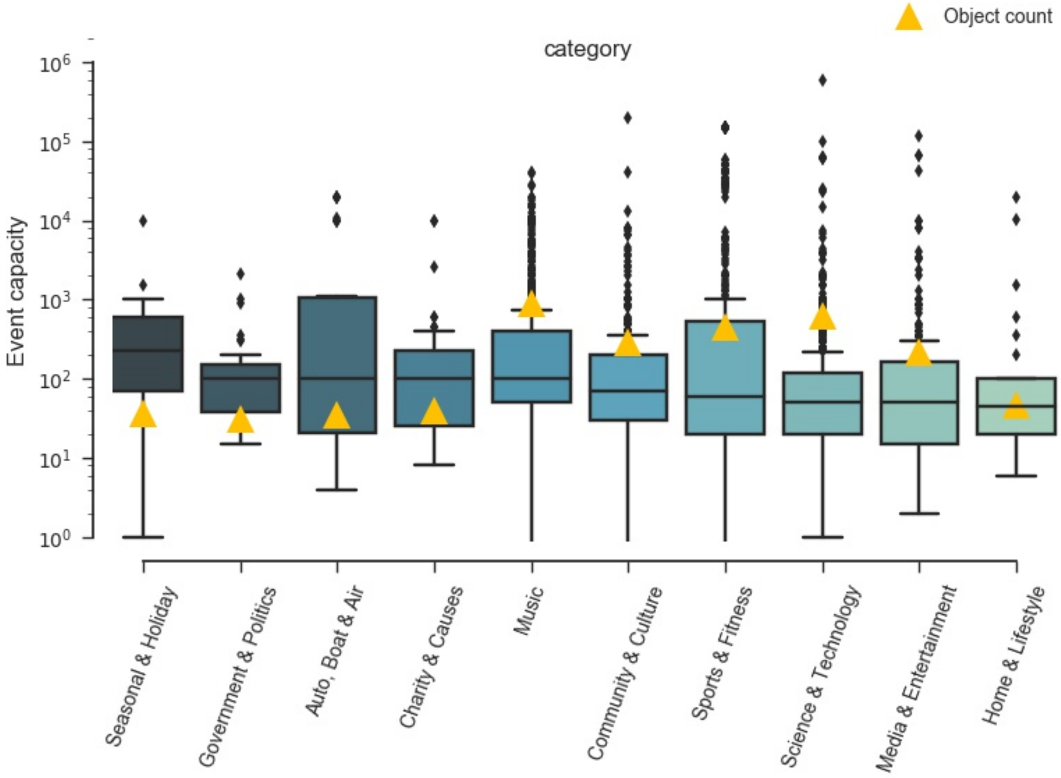


Figure 19. Eventbrite event capacity distribution

About 71% of them are paid while detailed price information is provided using a separate API endpoint. The dataset contains 20 different event categories. Figure 19 shows information for the ten categories with the highest median capacity as maximum number of attendees. This attribute is used as a popularity indication while no explicit attendance information, for example regarding ticket sales, is publicly available. It can be seen that the dataset covers a varied choice of event themes. Objects in the category ’seasonal & holiday’ have the highest median capacity while music events show the highest number of occurrences. Among all categories, music events are the number two most represented type. The vast majority of objects, however, is provided by business-related events. These account for 27% of the entire dataset but are not listed in Figure 19 as the median capacity only amounts to 28.5 users per event.

4.4.3 Ticketmaster

The Ticketmaster *Discovery API* (v.2.0) is publicly available and offers information for paid events that are marketed over the platform. It accepts standard GET requests and delivers JSON responses. The API offers convenient filter options that facilitate collecting event objects on a per-country basis. The API only includes events that fall into the current sales time frame. Thus, past events and objects that will take place in the far future cannot be retrieved [160]. At the test time, the API supplied about 1,900 events currently on sale for Germany in total. All of these included a specific start time while generally, no specified start time is required by the platform. Events can also be marketed on Ticketmaster if the exact start time is still to be determined. However, none of the obtained event objects made use of this option. Furthermore, none of the collected event representations included end times.

The dataset contains two categorical levels. Six superior categories comprise 44 subsegments that denote specific event genres. A clear tendency towards events in the entertainment sector is noted. The dataset does not contain business-related events. Figure 20 shows the ten most represented subsegments with their corresponding ticket price characteristics. For marketing purposes, a price range is provided instead of fixed values. Rock music and theater events are the most highly represented subsegments in the selection. A strong price variance in some segments, for example metal music, serves as contrast to other categories such as fine arts or circus events with low price variance. The dataset does not contain information that allows drawing conclusions on the event popularity. The number of tickets sold and other sales-related information is only accessible by the respective event organizers using a private API.

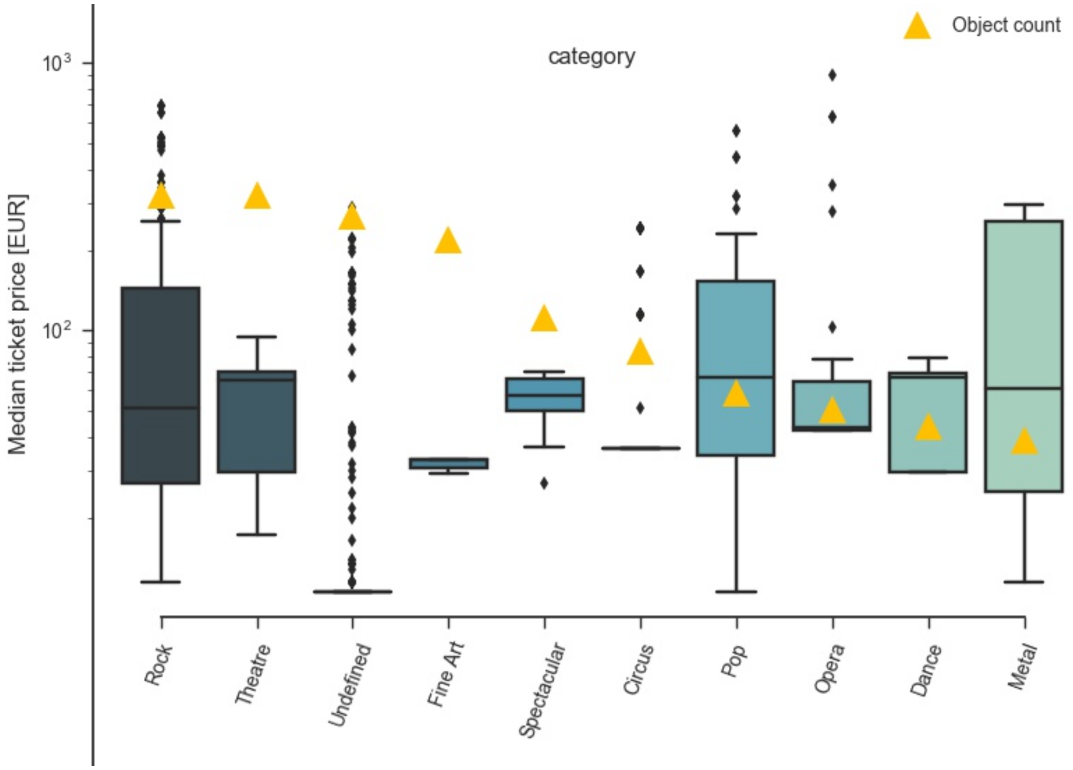


Figure 20. Ticketmaster median ticket prices for the most represented event categories

4.5 Summary of data collected

First, leading online data sources that generally provide information with regards to POIs and public events are preliminarily evaluated. Subsequently, algorithms are developed for scraping large amounts of data from the online platforms

Facebook, OSM, Eventbrite and Ticketmaster. The feasibility of scalable data collection as a basic prerequisite of feature integration from these sources is proven. Subsequently, the collected information undergoes an exploratory analysis to improve the understanding of data quality and its thematic representativeness.

CHAPTER 5

Data source benchmark

The presented data sources use various semantics and provide different sets of attributes. Facebook is the only considered source that provides public information that explicitly relates to the popularity of POIs and events. Other than that, the collected objects only contain indirect popularity information and typically only indicate the sole existence and themes of the event. This information is considered to be sufficient to derive certain feature subsets that have value for parking prediction models. For example, all data sources contain categorical information that allows drawing conclusions on thematically similar parts of the datasets. However, due to the heterogeneity of the data, merging of the acquired sources to a new, unified database is possible only to a limited extent. Missing popularity information for the benchmark sources cannot be reliably estimated from the available attributes. A unified database would be limited to basic information supplied by all integrated data sources. For this reason, benchmarking is primarily conducted to specify the extent and value-added by social media data compared to the alternative sources.

5.1 Duplicate identification techniques

It is assumed that especially highly popular events are reflected in different data sources. Higher popularity increases the chance of observing multiple heterogeneous representations. If unified databases are constructed and duplicate entries from the sources are not removed, the parking model input is ultimately biased and lacks real-life representation. Thus, duplicate objects need to be identified as an important part of the data preprocessing phase. For benchmarking purposes, duplicate identification is also relevant to clarify the number of exclusively supplied objects by a certain data source.

Zhang 2015 [161] developed a procedure for identification of both syntactic and semantic similarities among events from different data sources. In this study, a methodology is applied that focuses only on syntactic similarity but also considers categorical object matches with specific focus on duplicate identification. As the crucial part of object names rarely has a specific meaning, including a procedure focused on semantic similarity is assumed not to be beneficial. The following list provides an overview on all techniques developed for duplicate identification purposes:

1. Context matching: Geographic proximity, time similarity
2. Name matching: Similarity of name strings (syntactic)
3. Categorical matching: Similarity of object themes

5.1.1 Context matching

The first phase, *context matching*, is used to limit the scope of duplicate identification to a geographic focus area. This limits the amount of computational resources needed by reducing the number of objects to be processed in further steps. The size of the considered geographic area is chosen as a compromise of computational expensiveness and accuracy of location information in the available data. A square area with an edge length of one kilometer is chosen to account for inaccuracies of geographic references among databases.

5.1.2 Name matching

During the second step, *name matching*, the similarity of name strings from different objects is analyzed. Recchia and Louwerse 2013 [162] analyze the performance of 21 different algorithms specifically with regards to the similarity of place names. It was found that all approaches perform very differently depending on the language of the place names to which they are applied. For data taken form a German context, the *longest common substring* method was found to deliver acceptable results. Using the Python *difflib* (v.2.1) implementation of Ratcliff and Obershelp's *Gestalt Pattern Matching* algorithm [163], the longest matching string sequence is identified among two object names. The implementation extends the original algorithm by removing characters of low discriminant power such as whitespaces or blank lines. This procedure is recursively repeated for the remaining substrings. Finally, a similarity ratio is used to decide whether a name pair represents a match.

This approach works well to evaluate the semantic similarity of single words. For object names that are combinations of multiple terms, the technique's direct performance is rather limited. Considering the example POI name 'Oranienburg Tiergarten' (engl. Oranienburg Zoo), only a comparably low similarity ratio for the same content is calculated if the order of words is changed. In the literature, a popular measure to deal with this syntactic variation is the *Jaccard index.* It is used to calculate the similarity of two token sets derived from the names of the compared objects. The more common terms are found in both sets, the higher the Jaccard index. However, this measure only accounts for exact token matches. In case there are slight differences in name spelling, the algorithm does not perform well. For this reason, an extension is introduced that combines both the longest common substring method and the Jaccard index. The metric is calculated as presented in Formula 3.

$$M(A,B) = \frac{|C|}{\frac{1}{2}(|A|+|B|)} \,\Bigg|\, C = \{sim(a,b) > thres\} \text{ for } a \in A, b \in B \tag{3}$$

The token sets A and B refer to the respective object names. The set C is generated by calculating the longest common substring ratio (sim) for all possible combinations of A and B. If the obtained value for a pair of tokens lies above a certain threshold, the combination is counted as a match. This allows a certain variance in spelling of the object names depending on the threshold chosen. The ratio of matches identified divided by the average length of name token sets represents the final similarity metric M. Compared to the original Jaccard index divisor $A \cup B$, using the average token set length increases the metric's sensitivity in cases where object names show a significant difference in length.

For further accuracy improvements regarding identification of duplicate object names, filtering of token sets is conducted before the accuracy metric is calculated. This includes removal of duplicate tokens and stop words within each of the compared name sets. The stop word sets are specifically defined in accordance with the respective datasets.

5.1.3 Categorical matching

Matching of categorical object information is a primarily manual process. Each category of one data source has to be compared to all categories of the remaining data sources to identify one or more thematic matches. However, for the matching of OSM and Facebook POI categories, a full pairwise comparison is beyond the scope of manual assignment. With 143 OSM tags and 1,124 different Facebook categories considered, about 161,000 comparisons are theoretically necessary to identify objects in the same thematic space. For this reason, a heuristic procedure is developed that facilitates the assignment process.

After the syntactic preparation of the OSM tags to create a homogeneous format for category strings, the presented name matching algorithms are applied to find syntactic similarities. They are used to limit the amount of potential thematic matches to the five most promising candidates for each OSM tag. Subsequently, the prepared selection is corrected manually. This procedure delivers plausible semantic matches from syntactic similarity for about 60% of the OSM tags. Furthermore, a secondary procedure is developed that retrieves synonyms for name tokens of both sources from *WordNet*, a popular lexical database for English language [164]. These synonyms are used to add a semantic level to the categorical matching approach. Other than object names, tags and categories are combinations of general vocabulary and do generally have a semantic meaning. The entire set of token synonyms from one data source is compared to the category tokens of the second data source. Categorical matches are identified based on the syntactic similarity of synonym and alternative token using the extended Jaccard index metric. This procedures results in an additional four percent of the OSM tags to match with respective Facebook categories. All proposed category matches are manually verified. A share of further 13% of categorical matches is triggered by fully manual assignment. This process is supported by a self-developed tool to locate meaningful substrings in extensive string lists. The remaining 23% of OSM tags do not have any Facebook counterpart and are considered to be unique. In total, 300 out of 1,124 distinct Facebook categories can be matched to their corresponding OSM tags. This represents about 27% of all categories and about 20% of the entire dataset. A sample for POI matches regarding OSM and Facebook is displayed in Appendix F3.

Categorical matching for the considered databases on the event side requires less efforts due to the limited number of distinct categories for each data source. As manual assignment provides the best results, this approach is extensively applied for the given datasets and the assignment results are presented in Appendix T2. For the case of Eventbrite, seven categories cannot be matched with Facebook equivalents while 11% of the latter do not have a direct counterpart in return. For Ticketmaster, only nine matches are identified while one category remains without equivalent on the Facebook side. In fact, 76% of the Facebook categories cannot be matched with this data source. All in all, 23% of the Facebook categories have equivalents for both benchmarked sources while 11% are not reflected. The rest only matches with categories from Eventbrite.

5.2 Benchmark results

5.2.1 POI data sources

A supervised learning machine is used to automate the decision whether a pair of POI objects from different data sources represents a duplicate. All previously calculated matching indicators regarding object names and category are applied as input features for the classifier. First, 575 random samples from the Facebook dataset are taken and potential match candidates in the OSM dataset are identified using the described geographic context matching technique. For the specific sample set, this results in about 10,200 potential duplicates. Within this collection, all actual duplicates are manually identified and labeled to create a ground truth of object matches. Subsequently, the dataset is balanced by removing entities of the majority class 'no duplicate'. This is necessary to avoid biased classifier results and unrealistic accuracy scores. As the minority class 'duplicate' only represents 157 samples, the balanced labeled dataset comprises only 314 matches in total. More samples can only be acquired under significant further human efforts. Thus, a learning machine must be chosen that approximates the general trend well with small sample sizes. A decision tree model is used while hyperparameters are chosen that avoid overfitting. This includes limiting the maximum tree depth to three levels and a minimum required sample size of 20 for branch splits. If depth and split limits are not determined, decision trees tend to represent single objects instead of generalized patterns, ultimately leading to low performance.

When applying a three-fold cross-validation on the trained model, a classification accuracy of 97% is achieved while almost all of the perceived variance in the data can be explained with the previously introduced, combined similarity metric. As a second important feature, the Jaccard index contributes to the classification results while the longest common substring method and categorical matching are rather not relevant. The main reason for this behavior relates to the fact that 72% of all POI object names in the OSM dataset consist of more than one token. For the Facebook dataset, this ratio is about 87%. In fact, the original Jaccard index and the combined similarity technique are crucial for the success of the duplicate identification step. Figure 21 shows the two important features as axes of a scatter plot visualizing the dataset. Duplicates and real unique values are marked in different colors. A *jitter*, deliberate value variation, is introduced to facilitate the data point visualization. Without the jitter, data points frequently overlap and valuable information with regards to the extent in certain areas gets lost in the visualization process. To enhance the model understanding and fit, the split points determined in the decision tree generation are highlighted with grey rectangles. All samples that lie within these areas are predicted to be 'no duplicate'. It can be seen that most data points in the 'unique' class are located close to zero for both relevant features while the complementary class is widely distributed. A visual correlation between both indices is observed.

Applying this model to the entire POI datasets, about twelve percent of all OSM objects are identified as duplicates of Facebook POIs. This corresponds to roughly 267,000 POIs. 79% of these duplicates represent unique matches between the two data sources while 21% correspond to OSM objects that have multiple related Facebook objects. This phenomenon is typically observed if one object has multiple counterparts of inferior hierarchical level. For example, this can be shops or areas within large building complexes as object names for these are typically

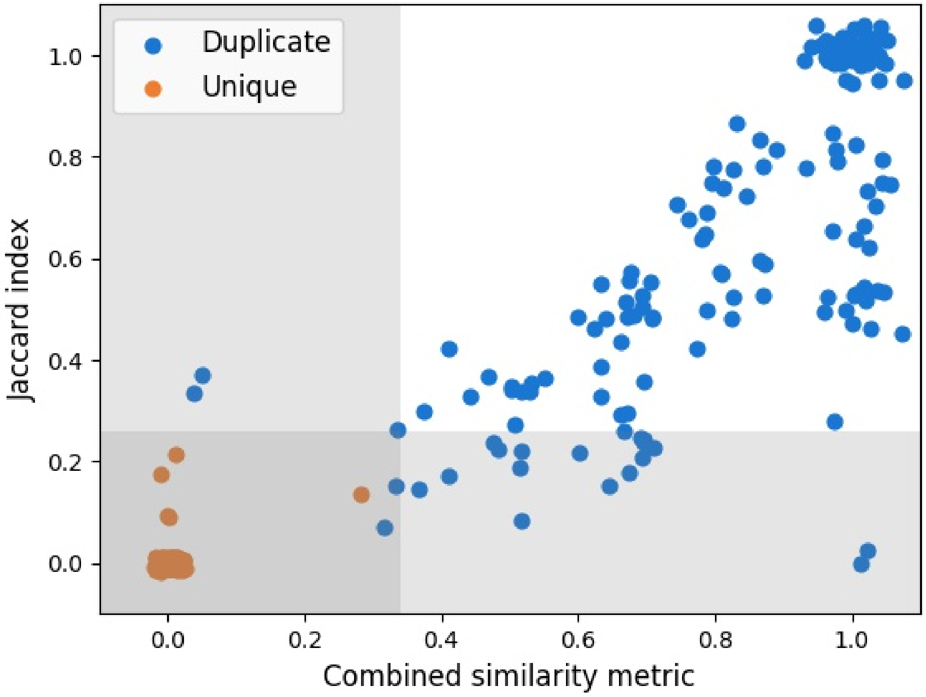


Figure 21. Labeled dataset for duplicate identification

very similar. About seven percent of all OSM objects have more than two matches.

Due to the feature heterogeneity of the analyzed data sources, full convertibility of objects from either source cannot be achieved. Ultimately, remaining entities from the original sources after removing duplicates can only account for distinct input parameters of the parking prediction model. In this context, Facebook is used as a primary data source due to the availability of object-specific popularity information. Categorical estimation of popularity cannot lead to similarly accurate, object-specific popularity attributes as a strong variability within all categories is observed (Chapter 4.2.2).

Furthermore, it is necessary to identify internal duplicates within the Facebook dataset. As these crowdsourced contents are not editorially managed, there is a high probability for users to contribute duplicate information. In the entire dataset, there are about 6,100 explicit name duplicates with differences regarding their identification numbers. To identify duplicates with non-explicit name matching, a labeled dataset with 375 entries is generated and different classifiers are compared using three-fold cross validation. It is found that a Support Vector Classifier achieves the best performance with about 91% accuracy. Regarding feature importances, the Jaccard index and the categorical similarity measure are primarily taken into account. The classifier's consistency among the single fold results represents the highest among all tested models with a standard deviation of only 3% for the accuracy measure. The benchmark decision tree classifier results in 15% standard deviation of the accuracy measure. Applying the classification to the entire dataset, about 64,800 duplicates are identified. Figure 22 summarizes the POI benchmark results.

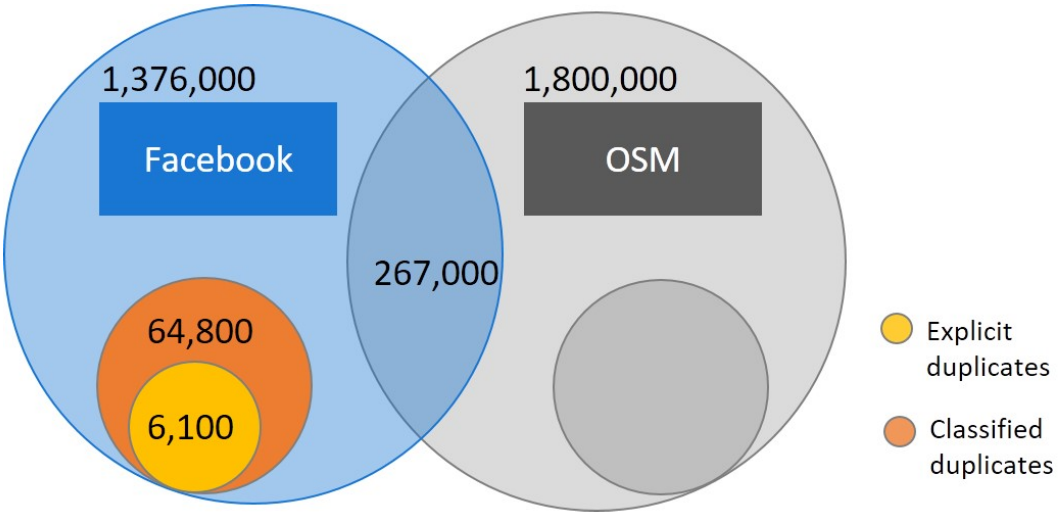


Figure 22. POI duplicate identification results

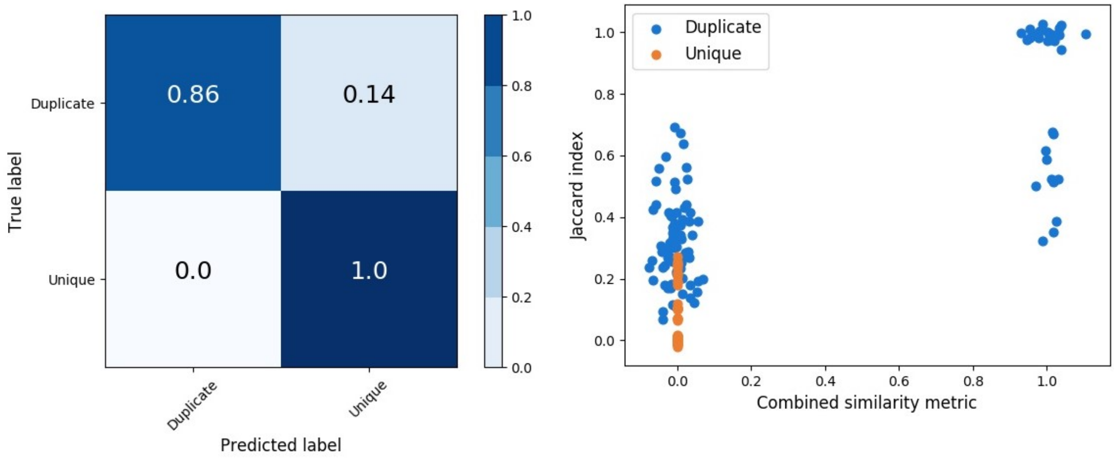
5.2.2 Event data sources

For duplicate detection in the event databases, contextual matching is extended by checks for similar time frames. First, the benchmark data sources are transformed from ISO 8601 timestamps into the UNIX system (e.g. 2017-

05-17T00:58:31Z -> 1494982711). A tolerance of two hours before and after the noted start time in the primary data source is introduced to avoid false exclusion of duplicates if there are deviations of the stated time. Simultaneously, a reasonable tolerance time frame reduces the number of potential misclassifications. Similar to the previous section, a labeled dataset is manually generated. As the number of objects in the benchmark datasets is very limited and time restrictions greatly reduce the number of potential matches, the full dataset is manually labeled for both, Eventbrite and Ticketmaster data. Even though similar features are used, the inherent data structure is particularly different to the POI matching problem. In fact, every relevant combination of data sources must be processed via training a separate learning machine.

For Ticketmaster data, the decision tree classifier only performs with an accuracy of about 77% based on a labeled and balanced set with 550 entries. Figure 23(b) visualizes the reason for this insufficient performance. A decision tree can only separate classes linearly. However, with the given features, an overlap of duplicate and unique entries can be recognized with regards to the considered attributes. The developed similarity metric is shown to avoid false positive (duplicate) reliably but cannot separate false negatives (unique). Thus, the decision tree assumes the Jaccard index to be the only important feature. Other models are trained and tested using three-fold cross-validation. A support vector classifier with linear kernel achieves 81% accuracy with optimized hyperparameters. Finally, a perceptron model is trained and 90% accuracy is achieved, representing the best value. Figure 23(a) shows the confusion matrix for this classifier. It can be seen that the false negative ratio is reduced to 14%. The perceptron considered both the original and extended Jaccard index as input features.

To compare Eventbrite and Facebook events, a labeled dataset with 635 entries



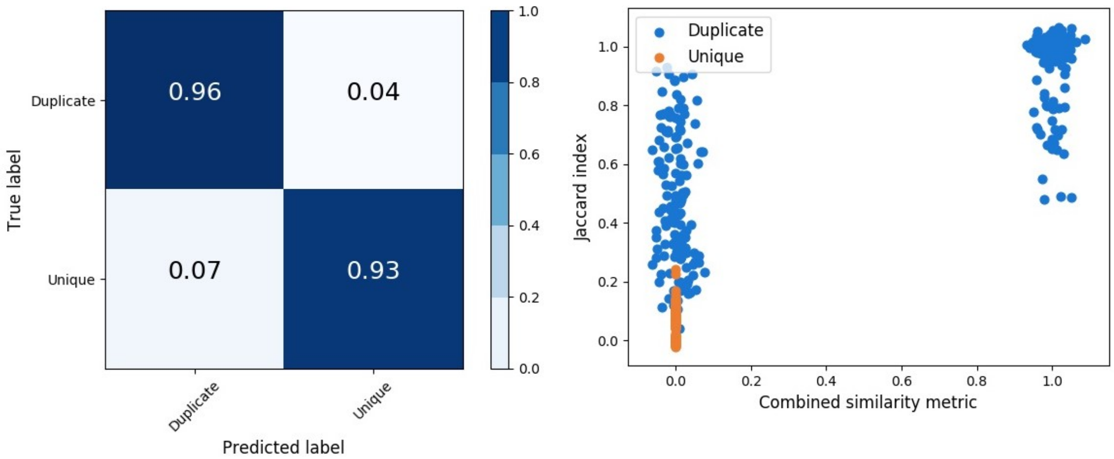
(a) Confusion Matrix (perceptron model)

(b) Data distribution

Figure 23. Labeled Ticketmaster event matches

is manually generated. A decision tree classifier is trained on a balanced subset with 560 entries. It achieves an accuracy score of 95% using three-fold cross-validation. SVMs and perceptron models achieve accuracies in the same range. Thus, the modeling approach with the easiest architecture, the decision tree, is selected. The original Jaccard index is observed to be the most important feature. Figure 24(a) shows a balanced distribution of false positive and false negative predicted labels. Even though the scatterplot in Figure 24(b) shows a similar pattern, the class overlap with regards to the Jaccard index is much smaller. This is the main reason for the observed high decision tree performance. It has to be noted that there are no duplicate events when Eventbrite and Ticketmaster are compared as both platforms are used as exclusive marketplaces. When limiting the Facebook dataset to the same time period covered by the benchmark data sources, about 59,600 event objects are to be considered.

Considering internal duplicates within the Facebook database, the subset considered for benchmarking purposes comprises about 4,100 unique contextual matches. About 350 of these are explicit name duplicates and just show different



(a) Confusion Matrix (Decision Tree)

(b) Data distribution

Figure 24. Labeled Eventbrite event matches

identification numbers. Matches with identical identification numbers are not considered as they represent comparisons of identical objects. A labeled dataset with 1,400 entries is generated and a decision tree classifier for duplicate identification is trained. Using the Jaccard index as single main feature, the model achieves 92% accuracy based on a three-fold cross-validation. The classifier is used to identify 650 duplicates on the considered dataset while all explicit object duplicates are also identified. In fact, the explicit name duplicates are a subset of the ones identified by the classifier. This confirms its validity. Figure 25 graphically summarizes the event benchmarking results. Duplicate entries in the Facebook dataset are responsible for about one percent of all attendees and interested users.

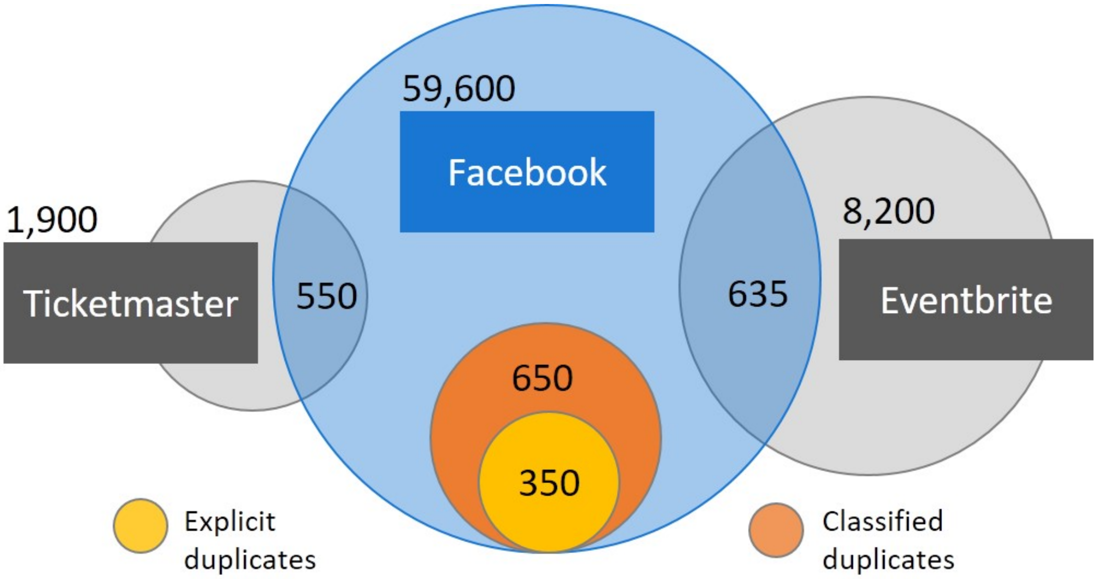


Figure 25. Event duplicate identification results

CHAPTER 6

Feature engineering

This chapter describes the developed procedures for transforming raw social media data into valuable sets of input features for parking demand modeling. This includes the adjustment of popularity attributes based on external resources, as well as the the text-based extraction of parking-related influences. Also, thematic similarities among POIs and events are analyzed with regards to their implications on automotive mobility. This includes cross-referencing Facebook data to historical parking events identified in FCD and feature extraction based on unsupervised toping modeling.

6.1 Feature extraction roadmap

To understand the role of social media data for modeling parking demand in cities, it is necessary to analyze the underlying relationship between the available and predictable information. For better modeling results, extracted input features must be as closely connected to the target attribute as possible. Figure 26 visualizes the general relationship between data in social media and the target measure parking occupancy. For each modeling stage, corresponding attribute sets are extracted to be used as input features.

Deriving parking occupancy information from social media data is a multi-stage modeling process. Social media users represent only a specific sample of the entire population with car access. As described in the literature, individuals go through a complex decision making procedure before using social media platforms. Also, strong differences among users are found regarding the general usage frequency and utilization of specific functionalities. In fact, the online popularity of real-life objects is seen as the result of a subsequent decision making process that

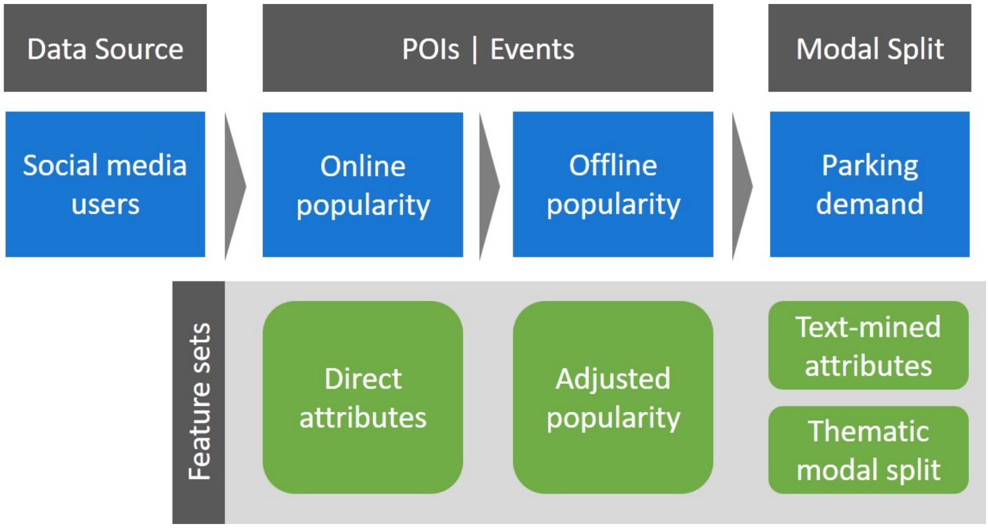


Figure 26. Interconnection between social media data and parking demand with extracted feature sets

involves particular individual contributions and interests. The users that interact with Facebook POIs and events are another subsample of the entire user base while their behavior is driven by personal motives. In fact, the directly supplied attributes reflecting the online popularity lack general representativeness.

The actual physical visitors of POIs or offline attendees of public events represent another subgroup of the users that interact online. As interaction in social media is not binding, offline attendance rates typically strongly differ from the online popularity observed. Using the concept of *adjusted popularity*, offline popularity measures are derived from the available online indicators. This includes referencing geospatial social media data to alternative, more general data sources. Also, findings from literature regarding offline popularity of online entities are taken into account. The adjusted popularity values are subsequently used as input features for the parking demand model.

The target measure, parking demand at a specific location, is directly influenced by the popularity of POIs and events in the area. Based on general data regarding the modal split preference, a subgroup of offline attendees choses the car as travel mode. This is highly influenced by certain characteristics of the online entity that are not provided in the form of structured attributes. To be able to extract this information specifically for attributes having an impact on travel mode choice, text mining is applied.

Moreover, it is expected that POIs and events with similar themes have similar effects on local traffic. For this reason, all online entities with sufficient textual information are analyzed with regards to their thematic orientation using different text mining techniques. This includes supervised as well as unsupervised machine learning methods. Subsequently, the themes are connected to historical FCD parking events to define a content-specific modal split. In combination with the available popularity information, this serves as feature set for parking occupancy modeling.

Generally, all levels in the modeling chain represent subgroups of the preceding stage. However, the actual measure for each level is also influenced by an unknown external amount of attendees or drivers that is not reflected by the POI or event data. In fact, all extracted feature sets are naturally just indicators and cannot reflect the actual driver behavior in an isolated manner.

6.2 Adjusted popularity measures

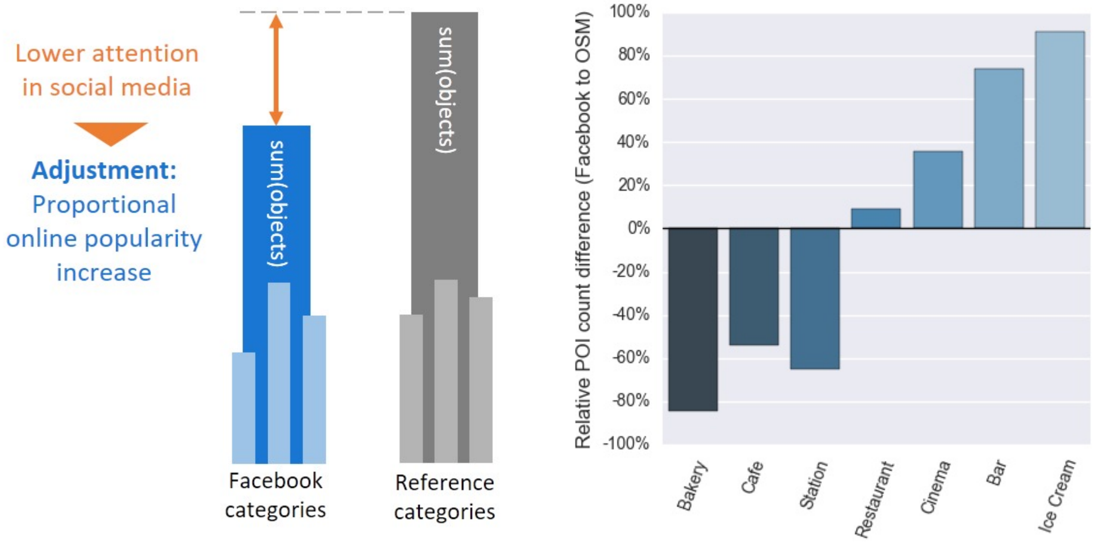
As described in the literature, physical visits of POIs and events depend on a complex set of factors related to the individual and social group attending as well as contextual and entity-specific factors. Due to a lack of personal user data, many of these relevant factors cannot be determined. For reasons of privacy protection, Facebook does not publicly offer this information. Thus, estimates regarding offline popularity must be based on contents that are publicly accessible and not privacy-critical at the same time.

The concept of *adjusted popularity* aims at correcting the bias that is introduced by the crowdsourced character of social media. As can be seen in the literature, thematising the representativeness of social media is determined by the over- and underrepresentation of certain social groups. In fact, the user-generated contents offered on these platforms represent only the interests of certain parts of society.

6.2.1 Adjustment using a reference data source

As OSM is a general mapping resource, it is assumed that the thematic distribution of its place objects corresponds to the actual real life occurrences. Based on a high number of contributors especially in Germany, the information contained is expected to be unbiased and highly detailed. Facebook, on the other hand, maintains a widely uncontrolled POI database without public review or correction processes. Thus, the over- and underrepresentation of the Facebook dataset is examined based on the thematic differences between both datasets.

First, city center coordinates for the 70 largest cities in Germany are used as a basis for creating quadratic polygons that define the focus area. As parking is an issue mainly in urban contexts, the representativeness of this section is assumed to be higher than the results when all of Germany is taken into account. Moreover, while OSM data is available on a nationwide basis, the data acquisition applied for the Facebook dataset limits its availability to urban areas. A nationwide comparison would lead to biased results that are avoided by focusing on city areas only. For the quadratic polygon size, an edge length of 25 km is chosen to achieve a sufficiently large subset of OSM and Facebook POIs for these areas. Taking into account the categorical matches defined in the course of the data source benchmarking (Appendix F3), the sum of thematically similar objects over all city polygons is calculated for both data sources. Subsequently, the relative difference between both object counts is calculated and used as a linear measure for adjusting the raw online popularity values. As exemplarily displayed in Figure 27(a), if the sum of Facebook objects is lower then the corresponding value on the OSM side, it is assumed that this POI category is underrepresented in Facebook. In this case, the given popularity attributes must be increased to correct the observed bias.



(a) General adjustment principle

(b) Relative POI count differences by category

Figure 27. Adjusted popularity with reference data source

Figure 27(b) shows several exemplary OSM categories and their relation to corresponding Facebook object counts. While certain categories are almost equally represented, for example restaurants, the relative differences vary greatly for other categories. Bakeries, for instance, are found to be highly underrepresented in Facebook with about 84% less objects counted for the test area compared to OSM. Bars and ice cream places, to the contrary, are meaningfully more represented in Facebook than in OSM. Figure 28 shows the distribution of relative object occurrences using a histogram with ten bins. Thin lines at the bottom of the diagram indicate occurrences of single values. It can be seen that the section indicating higher object counts on the OSM side (left) is stronger represented than its counterpart. Extreme differences up to almost -100% are indicated. Transforming these values directly into adjustment factors would practically remove the popularity information. Thus, the general extent of applied adjustment is limited to account for only up to 50% of the observed popularity. To make an example, a bakery with 100 fans on Facebook qualifies for a 84% popularity increase. Limiting the extent of the adjustment applied, an adjusted popularity of 142 fans is calculated. This procedure is repeated similarly with check-ins and the number of people talking about the POI. As the weighted influence of the adjustment procedure is arbitrarily chosen, additional feature sets using 30% and 70% adjustment influence are generated.

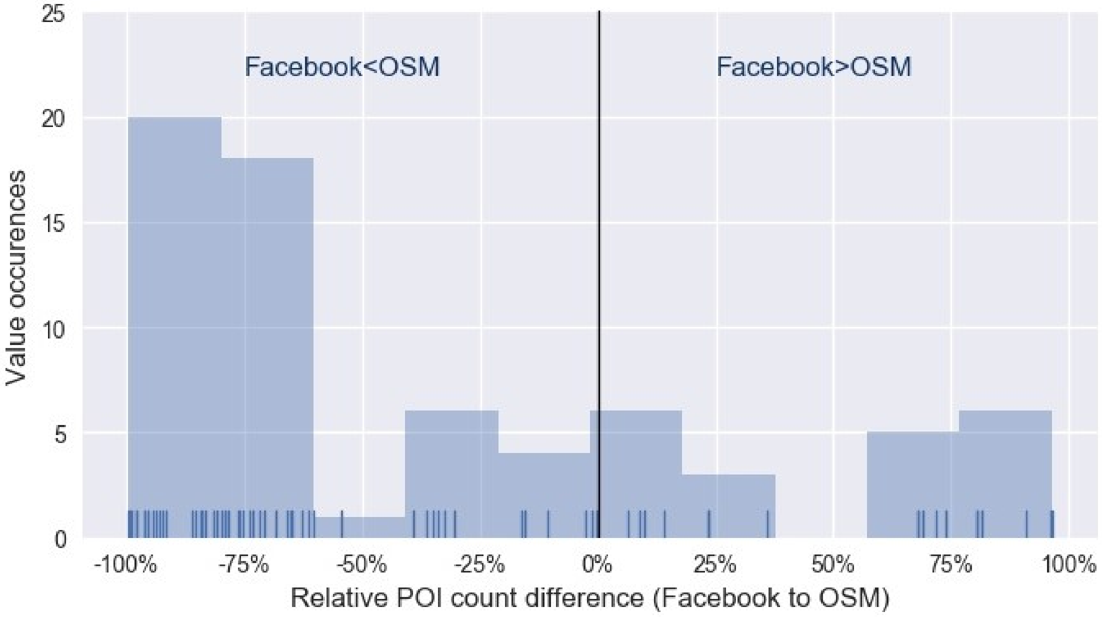


Figure 28. Facebook POI popularity adjustment factor distribution

Before the adjustment of online popularity can take place, categories with less than ten objects for any of the two data sources are removed from the analysis set. Small absolute differences among object counts for these samples would result in large popularity adjustment factors that do not reflect the general distribution. Taking into account this filter logic, only 69 OSM categories fulfill all described requirements. These correspond to 237 Facebook categories, representing about

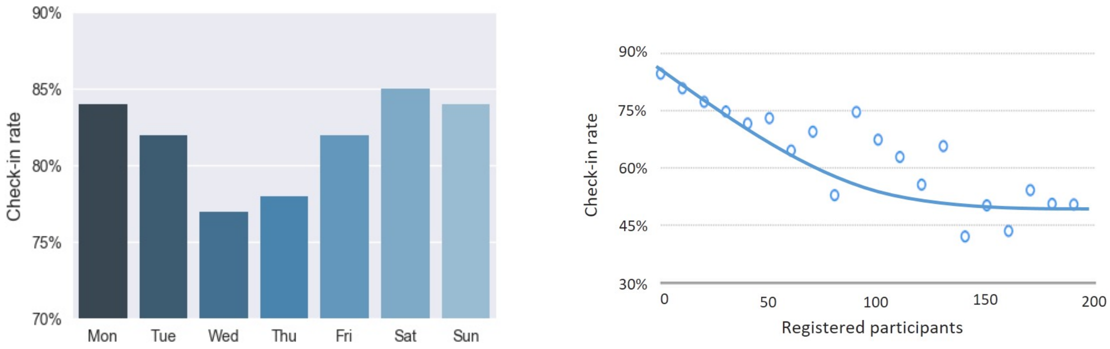
21% of all 1,124 categories. In total about 254,000 Facebook POIs are affected by the category-based popularity adjustment. This corresponds to about 19% of the filtered dataset. Popularity adjustment with OSM as reference data source cannot be conducted for the remaining POI objects.

For events, no generalized data source exists that reflects objects for various genres and target groups in an equal manner. The benchmark data sources taken into account, Ticketmaster and Eventbrite, are both considered biased based on their commercial focus. These platforms generally have a low interest in promoting free events as there is no direct revenue potential regarding their business models. In fact, popularity adjustment based on a reference data source equal to the procedure applied on the POI side is not possible.

6.2.2 Adjustment using domain knowledge

As there are no publications describing patterns or generalized influence factors for physical visits of POIs, findings from literature cannot be used to define adjusted popularity features. However, regarding public events, domain knowledge was published by the platform Doorkeeper [121] that is used as a basis for adjustment. As Doorkeeper is a Japanese event platform with a comparably small amount of offered events and a tendency towards professional themes, the data source describes a rather specific subgroup of all possible events. However, the information published provides interesting insights covering a general rather than an individual level. Domain knowledge with regards to the influences of weekday and event size is provided and displayed in Figure 29. Also, experiences related to event pricing are provided but cannot be accounted for in the context of popularity adjustment as Facebook does not provide corresponding data for cross-reference.

For each event in the Facebook dataset, the respective weekday is extracted as a separate attribute, being used as basis for matching the weekday-specific



(a) Check-in rate by weekday

(b) Check-in rate by event size

Figure 29. Event domain knowledge from Doorkeeper; based on [121]

attendance rates. Detailed information regarding the check-in rate depending on the number of event participants is only available in a graphical format. Thus, data points are manually copied and used to fit a trend line into the presented data sample. Size-dependent popularity is adjusted in accordance with this function. As only the value range up to 200 participants is presented, larger events are assumed to fall into the range of approximately 50% offline attendance. Other external data also reports confirming information for this range [120].

No domain knowledge is available with regards to the interaction of both considered factors, weekday and event size. Thus, the mean of both is calculated and applied to the popularity attributes in the Facebook event database. This represents an equally weighted relationship. Moreover, two additional feature sets are generated with directed weighting to each of the factors using a ratio of 70:30 respectively.

6.3 Text mining for feature extraction

As indicated in the literature, the influence factors determining physical POI and event visits, as well as travel mode choices are complex. Other than the considered popularity-related factors, no attributes are contained in the available data that carry explicitly relevant information. However, textual resources can provide implicit value to be extracted. Thus, different binary features are derived from unstructured text and used as input for parking demand modeling.

It has to be noted that the Facebook POI dataset does not contain text other than the object naming and categorical information. No place descriptions are available. Moreover, the explicitly relevant factors that determine parking demand are not sufficiently researched with regards to POIs. In fact, attribute extraction cannot be applied to the available POI data. To the contrary, the Facebook event dataset contains descriptions for each object that is formulated and published by the organizers. This is seen as a rich source of textual information to be used for feature extraction. In the literature, certain explicit factors have also been identified as relevant for event attendance and travel mode choice.

6.3.1 Target group attributes

Facebook does collect highly detailed target group demographics and behavioral information. However, for privacy reasons, this data is not publicly accessible. Instead, the public event information must be analyzed to determine attributes that describe the social group targeted by the event. First, a labeled dataset with 500 random event objects is manually generated to be analyzed using supervised learning methods. As text input, event names and descriptions are taken into account while the names and categories for the event location are also included. The categorical information for the event itself is not considered as it is only available for less than 5% of the dataset.

A binary attribute is introduced to identify events that attract mostly elderly people. As this demographic is generally found to prefer car usage over other modes of travel, events focusing on this group are expected to create a disproportionately high parking demand compared to other events with similar popularity. Secondly, as higher income also correlates with car usage, a label is introduced to identify events that attract mostly wealthy people. Lastly, a label is added to identify events that attract environmentally-aware people as these tend to avoid car usage and prefer public transport. Even though male individuals are more likely to use cars than females, it is hard to identify event contents that are gender-typical. Thus, this factor cannot be reflected in the analysis. Furthermore, it is expected that reflecting event pricing with separate labels is not beneficial. The online popularity of event objects is independent from more formal registrations in direct contact with the organizer. Thus, findings that indicate higher attendance rates for paid registration-only-events cannot be directly applied to the available data. It is assumed that attendance rates and travel mode choice are not influenced by variance in event pricing on a general level.

By calculating the share of objects that the above mentioned attributes apply to, it is found that each of them is relevant for less than 4.5% of the labeled dataset. For supervised classification tasks, this results in highly unbalanced datasets and biased accuracy measures. When balancing the labeled dataset by randomly removing samples from the majority classes, only a comparably small subset of the generated data can be used for machine learning. Using word-based features for the classification, there is only a very limited number of multiple occurrences for identical words among documents. As event descriptions use highly diverse language containing slang and special characters, this effect is intensified. Additionally, about 1.4% of all labeled events have descriptions in other languages than German even though the events are held in Germany. This also increases the total number of distinguished words in the dataset while multiple word occurrences are not affected. All in all, this leads to an extreme sparsity of the generated feature matrices for a small, balanced training set. If a learning machine is trained on this set, overfitting on the available data is observed. Thus, these target-group-related attributes cannot be used for classification. In fact, a high number of objects have to be labeled to improve the model generalization.

6.3.2 Event content attributes

Labels with higher penetration must be chosen to increase the theoretical generalization potential. Weather conditions are assumed to have a direct influence on the offline attendance of events that are held outdoors. Also, under certain conditions, the share of car usage in terms of modal split is influenced by the local weather conditions. Thus, a corresponding label is introduced while about 15% of the labeled dataset account for outdoor events. Moreover, a label is introduced that focuses on alcohol consumption during the event. As this is expected to decrease car usage among the attendees and promote car pooling, it is potentially relevant as feature for parking demand as target variable. By counting the number of positively labeled objects, it is found that 32% of the labeled events show alcohol involvement.

For about ten percent of the data, a human labeler cannot determine the outdoor and alcohol attribute solely based on the available text data. This can be based on a lack of text in general or a lack of significance towards the underlying event themes. As the quality characteristics of the available text depend on the organizer input, a varying informative value is observed. Thus, for each binary attribute, a separate target class ’unknown’ is introduced, that stands for cases of unclear classification. If the labeled dataset is balanced as previously described, for both considered attributes, a random subset of about 150 samples is obtained that contains equal numbers for each of the three classes.

Tf-idf is used to generate feature matrices from the tokenized text collections. For token filtering, a list of stop words is used that contains characteristic content for German and English language, city names, annual figures and other text regarded as non-valuable from a domain and corpus-specific perspective. Stemming is applied to prepare the remaining tokens, leading to 11,500 distinct terms considered. As a separate feature, the object-specific number of words in the available text data is added as a new feature. As a lack of information is the main reason for labeling as 'unknown', the amount of text available for classification being introduced as a separate feature is expected to improve the identification of positive samples. To fit the value range of the tf-idf features in order to avoid model confusion, the text length feature is standardized.

A set of nine different learning algorithms is tested in default configuration on the available data. As both, the outdoor and alcohol label show three distinct classes, separate models for each class are generated. All samples of this class are regarded as positive while all other labels are considered to be negative. The predicted label is based on the highest obtained confidence for a generated model [165]. This approach is known as *one-vs-rest* strategy [166]. The selection of tested models comprises two naive bayes classifiers with Gaussian and multinomial kernel functions, a SVM classifier with linear kernel, an ANN and a decision tree. In terms of ensemble learning algorithms, a random forest, a stochastic gradient descent classifier and a voting classifier using the linear SVM and random forest model as basis for majority voting decisions. Furthermore, a dummy classifier based on uniform guessing is implemented as reference for the model performances. Figure 30 shows the obtained classification accuracies for the outdoor attribute focusing on four models in relation to the number of tf-idf features considered.

The Naive Bayes model with multinomial kernel shows performances of up to 60% accuracy for relatively small numbers of considered features. The Gaussian kernel leads to similar performances with 120 features while the multinomial kernel decreases in accuracy with increasing feature count. The decision tree performance

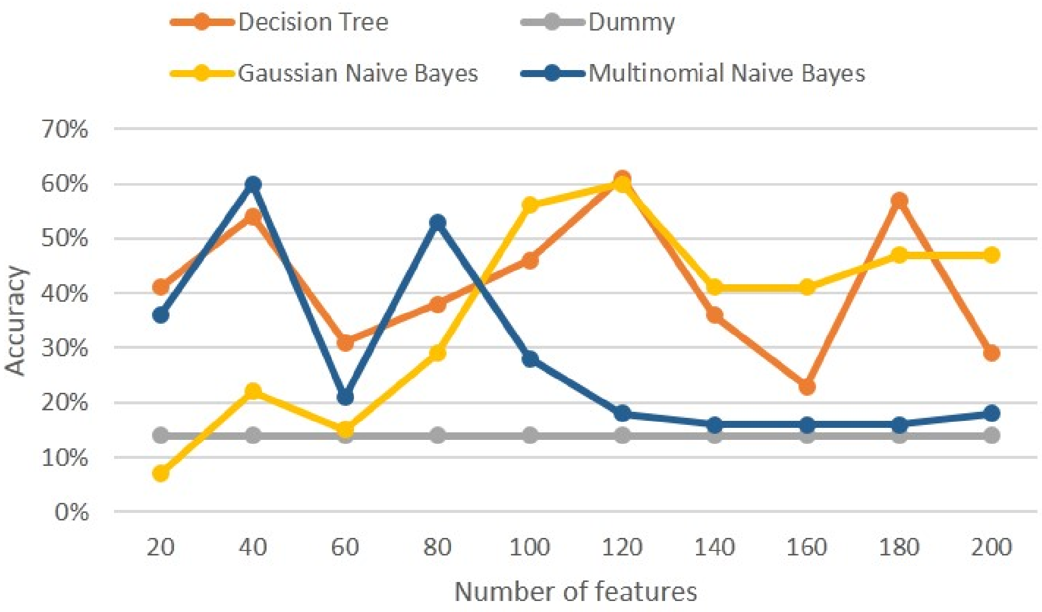


Figure 30. Classification accuracy for outdoor attribute

strongly varies with a peak at 120 features. The corresponding visualization for the alcohol attribute is displayed in Appendix F5.

Given the limited extent of the labeled dataset, optimization of the model hyperparameters leads to overfitting. This is confirmed by the high variance of the three-fold cross-validation scores obtained when using an exhaustive parameter grid search for model optimization independently form the chosen model. Even though higher model accuracies are achieved based on a one-fold comparison of prediction and test data, applying different training and test data leads to low observed model performances. Thus, no robust model improvements are achieved using parameter turning in this case. Considering confirmed robust configurations, the achieved performances are observed to be insufficient for further modeling applications. If preparatory models are used for feature extraction, the introduced error is directly implied on the actual target value - parking demand. As both, text mined attributes related to the event target group, as well as regarding specific event contents cannot be realized in sufficient quality, the attribute-specific modeling is considered infeasible.

6.3.3 Estimation of sample size required

As labeling requires significant manual resources, it is beneficial to estimate the number of samples needed before the actual labeling is conducted. This improves the planning capability of machine learning tasks and helps managing the necessary labor for labeling. Figueroa et al. 2012 [167] introduce an estimation methodology that is based on fitting a generic learning curve to empirical model performances based on varying sample sizes. Learning curves are generally found to follow inverse-power law functions [168]. Equation 4 shows the detailed relationship between curve parameters and obtained prediction accuracy. As the classifier increases asymptotically, a determines the minimum achievable error. The parameter b defines the learning rate while c sets the decay rate of the function [167].

$$\mathrm{Acc(x)} = (1 - a) - b * x^n \tag{4}$$

In the available literature, the number of features being taken into account is not specified. Thus, a flexible feature spectrum is introduced that represents about one percent of the distinct words in the sample set. A low number of samples interacting with a relatively high count of considered features leads to many irrelevant inputs being taken into account. Lowering the number of features in correspondence with the size of the corpus reduces classifier confusion.

The observed model accuracy highly depends on the chosen learning algorithm, influencing the sample size estimation. To balance this effect, all previously considered modeling approaches are applied on random subsets of the labeled dataset and the arithmetic mean of the obtained accuracy values is calculated to achieve partly independence from single classifiers. As the dummy model works independent from the sample size, it is excluded from the calculation. Figure 31 shows the mean accuracy in dependence of the considered number of labeled samples for both, the outdoor and alcohol attribute. Also, the number of features taken into account at each level is displayed with bars. Five sample size levels are defined for the analysis that lie lower or equal to the maximum available number of 150 balanced samples for each of the two considered attributes.

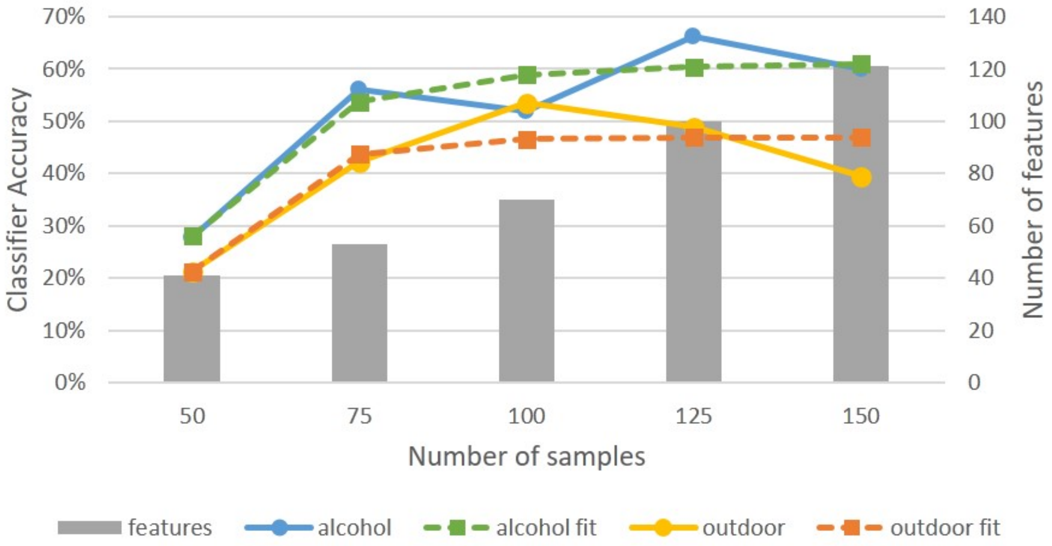


Figure 31. Classification accuracy by number of samples considered

For each target attribute, inverse-power law functions are fitted to the obtained values using a non-linear optimizer. The mean square error (MSE) between observed and fitted model accuracies is used as minimization target. For the alcohol attribute, a fitted 0.9% MSE is achieved. The outdoor attribute fit leads to an MSE of 1.1%. It can be seen, that in the range of 100 and more samples, the fitted curve shows an asymptotic behavior. In fact, an increased sample size does not lead to significant accuracy improvements. According to this analysis, the possibly achievable performance for the alcohol attribute is expected to be at about 61% while the outdoor value can be classified with a maximum of 47% accuracy. In fact, even with high resource contributions for labeling data, no major classifier improvement is expected. In contrast to the idealized learning curves in the literature, increased sample counts do not necessarily lead to accuracy improvements.

The actual benefit of additional samples for classification is highly dependent on the sample contents. If textual data is added that describes other concepts than the already considered samples, classifier confusion may be the result. The rapidly increasing number of distinct words within the considered sample range indicates diverging content being added. For large sample sizes, the rate of new terms being added is expected to decrease as the extent of potential new content must be part of the event domain. In fact, the analyzed sample sizes do potentially not represent the full pattern adequately as they cover only a small value range. Testing significantly higher sample counts may lead to different findings but this cannot be tested due to scope limitations of this study.

It has to be noted, that randomness is introduced by the algorithms that balance and limit the labeled dataset to a certain sample size. For both steps, samples are chosen at random to build up equal class distributions at the desired sample size. Thus, iterative testing with identical sample sizes may lead to diverging performances. Based on the observed heterogeneity of the corpus, data selection is an important reason for high variances of the observed classifier accuracy. To balance this effect and to get a robust performance estimator, each constellation of sample size level and target attribute is covered three times and the mean accuracy of all iterations is reported (Figure 31).

6.4 Thematic modal split modeling

6.4.1 Parking events as direct modal split indicator

Real-life entities in similar thematic areas are expected to show a strong resemblance regarding their implied mobility demands. Urban bakeries, for instance, typically experience peak occupancy in the mornings when people have breakfast. Differences among individual businesses are assumed to result uniquely from entity-specific factors such as marketing, product quality and geographic location. As bakeries are defined by a specific business model with certain customer interaction patterns, the travel mode choices of customers are assumed to follow specific distributions for all bakeries. To find features that account for these patterns, historical parking events identified in FCD are taken into account. As these are directly derived from a large number of vehicle-based trips over a certain time period, they provide an immediate indication of location-specific mobility conducted by car.

Parking events from October and November 2016 are taken as a sample of the available historical data. Matching procedures are applied to identify connections to Facebook objects in the 70 largest cities in Germany. Areas of nine square kilometers in each of the city centers are used as boundary box for comparison. The analysis is expanded to multiple cities to take different local settings into account. This is necessary as each city may have unique characteristics that have an impact on the parking behavior. Results for the entire set of areas are assumed to be more representative on a nationwide level. Figure 32 illustrates the methodology applied for extracting specific features based on parking events. First, historical parking occurrences are provided for the test area. It is not distinguished between successful and unsuccessful events, as well as between arriving and leaving trips. Subsequently, individual polygons with Facebook objects in their respective centers are constructed to define individual focus areas. Attributes of parking events and other objects within these areas are accumulated in variant forms for feature construction. These procedures are described in detail in the following sections.

It has to be noted that parking events are extracted from FCD using specific

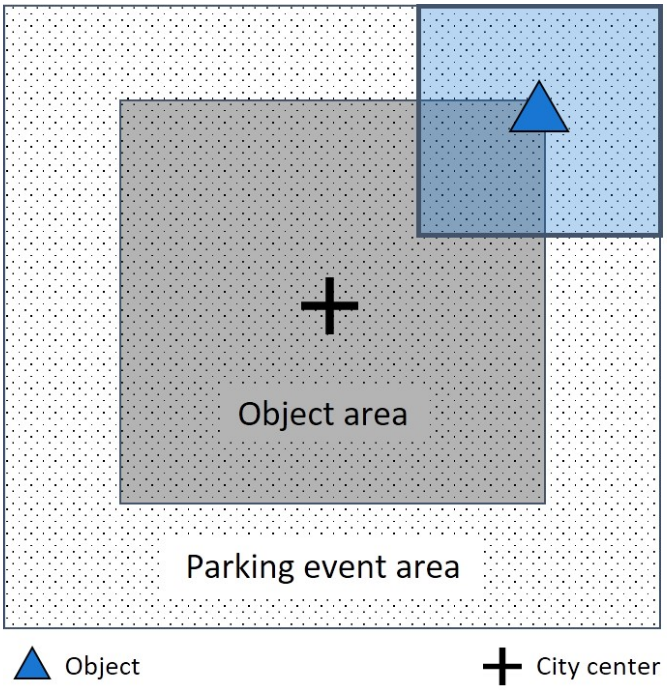


Figure 32. Cross-referencing parking events to Facebook objects

machine learning models. However, even though these instances are expected to represent the real-world parking demand very closely, they are certainly no exact representation and cannot be used as ground truth. First of all, only a source-specific subset of all vehicles supplies FCD and it is unclear whether this sample adequately represents the actual local mobility demand being satisfied by car. Secondly, as a model-based parking event extraction is applied, an identification error is introduced. No formal ground truth is available to estimate the relevancy of this type of error. Thus, parking events are seen as an indicator for the local parking situation rather than an explicit representation.

6.4.2 Category-specific parking demand for POIs

All obtained Facebook POIs contain category information based on the circumstance that this field is mandatory during data collection. With 1,124 different categories, the thematic separation is very detailed and parking events can directly be matched with the respective values. In total, two category-specific feature sets are extracted with regards to POIs and parking events. The first one focuses on parking demand observed during POI opening hours. The second one provides aggregated parking indicators on an hourly level.

Opening hours As most POIs are limited to certain opening hours, parking demand is only triggered by the object within these timeframes. Thus, opening hour information that corresponds to POIs in the analysis areas is used as a filter criterion for local parking events. Facebook provides opening hour fields for up to two shifts per day and parking events during these timeframes are summarized. Individual object polygons with 500 m edge length are chosen to define the potential influence area for the POI. This represents the expected walkable distance for POI visitors after having parked their cars.

The city center test areas cover POIs from 842 different categories. As a certain number of samples is required to derive meaningful patterns, all categories with less than five corresponding objects are neglected for further analysis. Furthermore, in average, about 70% of all considered objects do not comprise opening hour information. Thus, only objects from 287 different categories can be taken into account to determine parking events specifically related to the POI opening status. Figure 33 shows the number of accumulated parking events per weekday within the focus polygons around shopping malls. It is distinguished between parking events during and after the regular opening hours. 29 objects serve as basis for these findings while 80 further objects cannot be taken into account due to missing opening hour information.

The data indicates significant popularity of shopping malls on Saturdays while on Sundays, most objects are closed and the total parking demand is at its lowest point. Thursday also represents a weekday with low general popularity while the

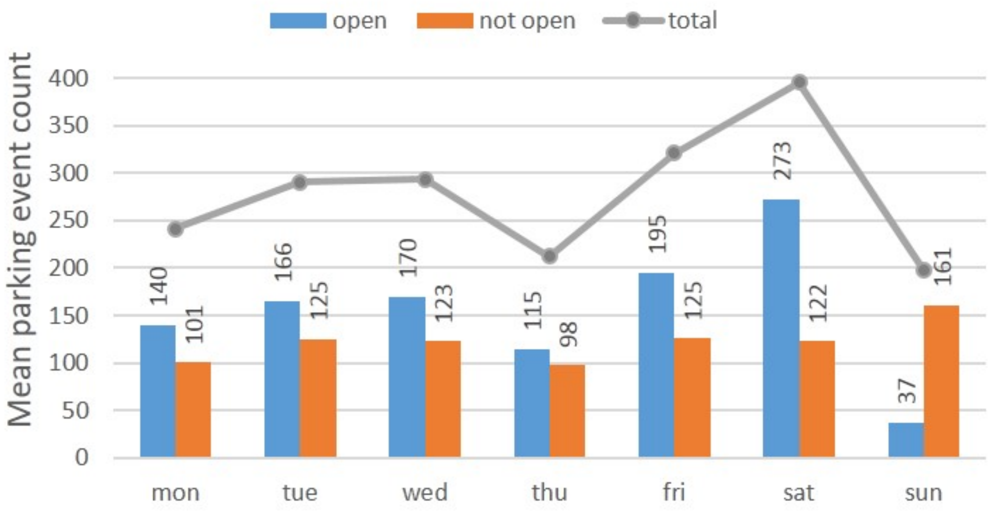


Figure 33. Parking events at shopping malls during and after opening hours per weekday

parking event count during opening hours is higher than the one corresponding to the remaining timeframe for all weekdays. This can be explained by the observation that the analyzed objects are open twelve or more hours per day, particularly during the active traffic periods. The considered facilities are closed primarily at nighttime when parking demand generally is low anyways. In fact, the effects of the POI cannot be perfectly separated from the underlying parking patterns using this approach.

Hourly analysis As initially described, POI-related mobility patterns have a strong connection to the time of day. Restricting the analysis to the binary differentiation between opening hours and other times is assumed to be too imprecise and may be misleading. Thus, a detailed aggregation of parking events per hour of day and weekday for each POI category is conducted. Figure 34 shows the exemplary results for shopping malls. A focus polygon with one kilometer edge length is used for data collection. The vertical axis indicates parking events per hour. Color coding on the plotted surface facilitate the differentiation between segments. While almost no parking events happen during the nighttime, strong peaks are observed for Saturdays around noontime. On the weekends, the number of parking events around midnight is higher than for the respective periods during the week. This is interpreted as a result of longer opening hours on the weekends and potentially special events. Compared to other weekdays, the parking demand on Thursdays is meaningfully lower than on other days. All provided parking event counts are mean values over all objects within this category. After all category-specific patterns are computed, each combination of hour and weekday is included as an individual feature.

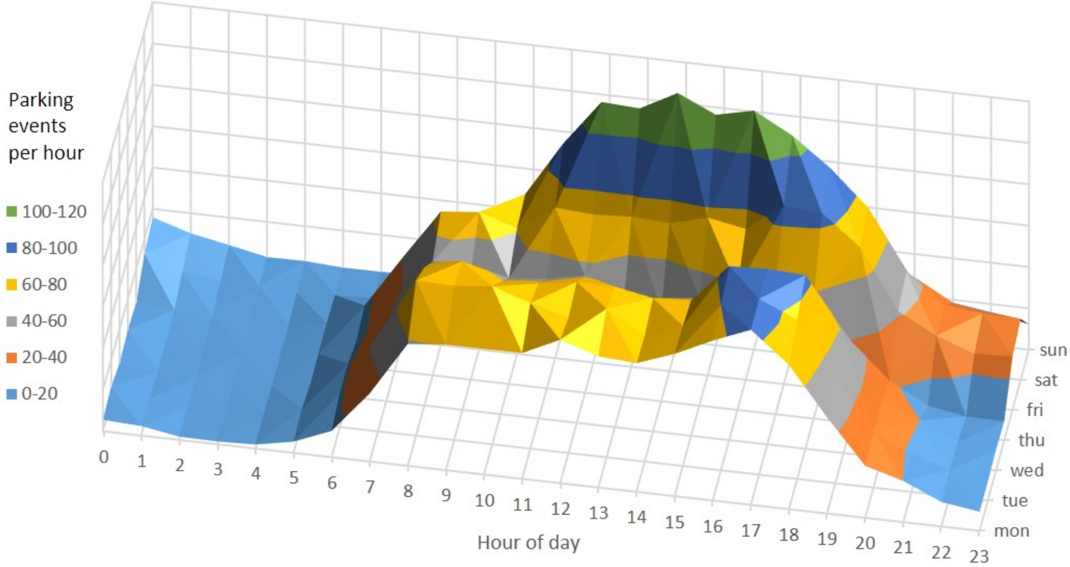


Figure 34. Parking events at shopping malls per hour and weekday

6.4.3 Category-specific parking demand for events

As event objects also comprise a category attribute, this can also be used as an indicator for thematic object similarity. In fact, transforming parking event patterns into features is equally possible for this data. However, only about 90,200 events have set category information while the remaining objects miss values for this attribute. In fact, more than 95% of the available data cannot be considered for this feature extraction method. Thus, preparatory machine learning models are generated to construct the missing category values to have a basis for thematic comparison among event objects.

Category predictors First, 40 original categories (Appendix F2) are used as basis for balancing the dataset. A 30% tolerance regarding the number of individual class samples is applied, leading to 3,200 labeled sample objects that can be used for analysis. The tolerance limit defines the allowed difference among class representations based on the number of objects in the balanced dataset. Higher values lead to disjunctive sets with generally higher number of objects contained. The crucial minority class for this case is the category 'Other' with only 61 samples, specifically denoting events that do not fit in the original attribute schema.

Using the set of learning algorithms from Chapter 6.3, tf-idf in combination with stemming and filtering for stop words is applied to the available text for feature preparation. Also, the word count of the respective object descriptions is added as standardized input. Figure 35 shows the highest precision among tested learning machines in default configuration for varying numbers of td-idf features considered.

Superior categorization As the obtained category models are not satisfactory with regards to their error implications, the chosen strategy has to be changed. As the reduced extent of the labeled dataset after balancing excludes the majority of available data, learning potential is lost. Thus, the detailed original category schema is replaced by the summarized system used in Appendix F2. The schema consists of eleven superior categories: Business, entertainment, fitness, food, music, nightlife, religion, shopping, volunteering, workshop and miscellaneous. Business events are the least represented with about 1,400 labeled objects. Applying a 30% tolerance in balancing, a labeled dataset with about 20,000 objects is created.

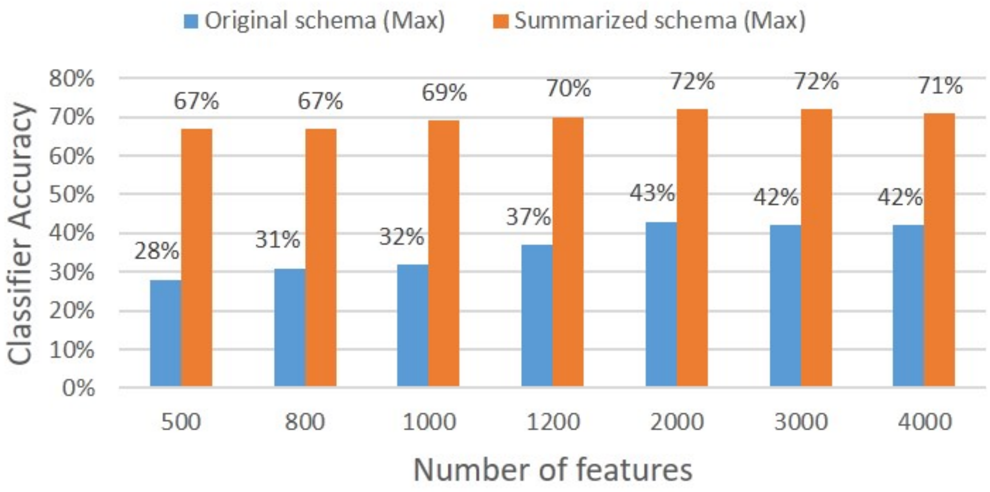


Figure 35. Classifier accuracies for event category reconstruction by number of features considered

Figure 35 shows a summary of the maximum classifier performance achieved for different feature sets. It can be seen that the summarization of categories leads to improved classification results. However, the observed trained models are still only at about 70% accuracy. This is insufficient for using the developed models as part of a preparatory process to define the missing event category values. It is assumed that an improved number of training samples may lead to better classifier performances. Required sample size estimation in accordance with the inverse-power law methodology indicate a 95% model accuracy at about 100,000 training samples. These cannot be obtained in the context of this thesis due to resource limitations. Thus, category-specific parking event features are not taken into account.

6.4.4 Parking demand from topic models

It was found in the previous section, that the category scheme supplied by Facebook or alternatively a summarized derivative cannot be applied as a basis for parking demand modeling. While a lack of usable training samples is expected to be the main reason, an alternative explanation for the observed model performances can be found with regards to the definition of the used classes. Real-life events are mixtures of multiple themes that can potentially interact. Binary decisions for or against a certain class label neglect this complex relationship and choose only one of many potential contents. Taking into account the previously proposed, summarized event categorization, certain inconsistencies of exclusive categorical labeling are observed. Music events, for instance, are likely to involve some sort of entertainment-related activities and potentially consumption of alcohol that is primarily covered with the category ’food’. Secondly, many category labels are imprecisely defined, particularly in the ’miscellaneous’ group. Events in the original categories ’community’,’meetup’ or ’neighborhood’ can account for a variety of activities that involve different individual behavior of attendees. Basically, these classes can contain events that cover arbitrary themes involving multiple persons. For these reasons, LDA as an unsupervised topic modeling approach is applied to avoid exclusive class separation and to introduce probabilistic thematic indicators.

Multiple LDA models are generated based on tf-idf features for the textual data available for 50,000 and 100,000 random samples of the Facebook event dataset, respectively. Event name, descriptions, place name and place category are considered. Even though the available event dataset is much larger, batch-based LDA training requires all samples to be loaded into memory. Given large sparse matrices with 4,000 tf-idf features, memory limitations of the available hardware determine the computational limits of the problem. Multiple LDA models are constructed for different numbers of topics to be identified in the dataset. For both training sets, a test set of 10,000 unseen event objects is transformed into adequate feature vectors and used to calculate the respective model perplexity. The obtained scores are shown in Figure 36.

It turns out that the simplest model with only five differentiated topics is

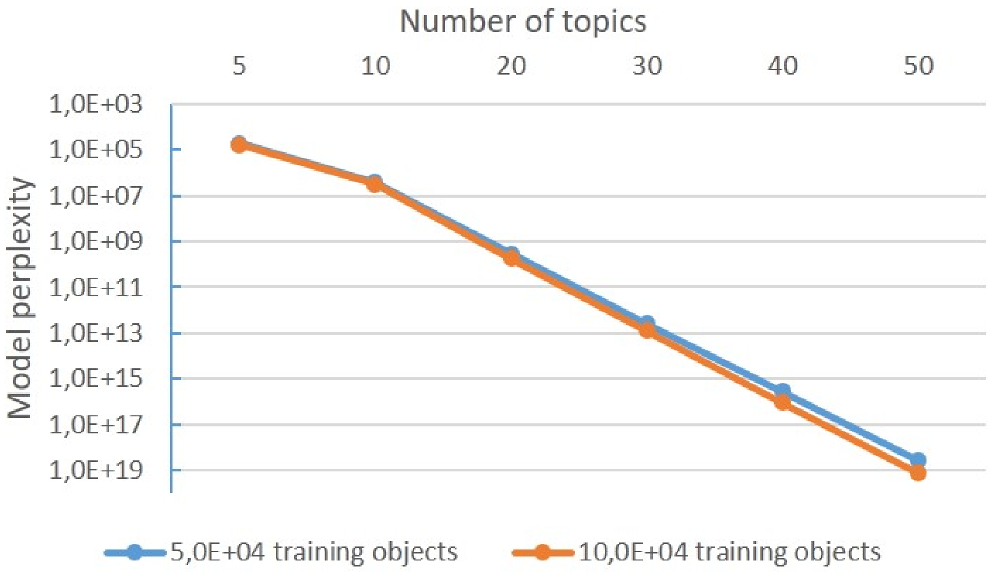


Figure 36. Obtained perplexity of LDA models by number of topics and training set

the most accurate one. Additionally, the perplexity values obtained are found to be almost independent from the amount of labeled training data used. Figure 37 visualizes the fit of the LDA model with five topics and the test corpus. Each circle represents one topic and its penetration of the corpus based on the share of tokens contained. A title is manually added for each topic as this step is not part of the topic modeling process. In accordance with the features being identified as relevant, the topic *Shows* comprises all events that involve performing live artists, movie showings or similar. *Sports* objects denote events that focus on rather passive watching of athletes playing sports games, races or similar. *Fitness* events, to the contrary, involve physical activity of the attendees. The topic *Shopping* relates primarily to marketing activities initiated by local businesses. The *Party* topic comprises clubbing and dancing events. The location of circles visualizes the thematic distance among identified topics. The input feature matrix is projected to the two-dimensional space using multidimensional scaling. This approach is beneficial for visualization as it emphasizes thematic distances, leading to clear topic separations. The topic *Party* is chosen as an example to highlight the relevant features being identified. As terms are stemmed in a preparatory process, the relevant features represent only partly actual words. For all features, the overall term frequency in the corpus, as well as the estimated frequency within the exemplary topic, are calculated.

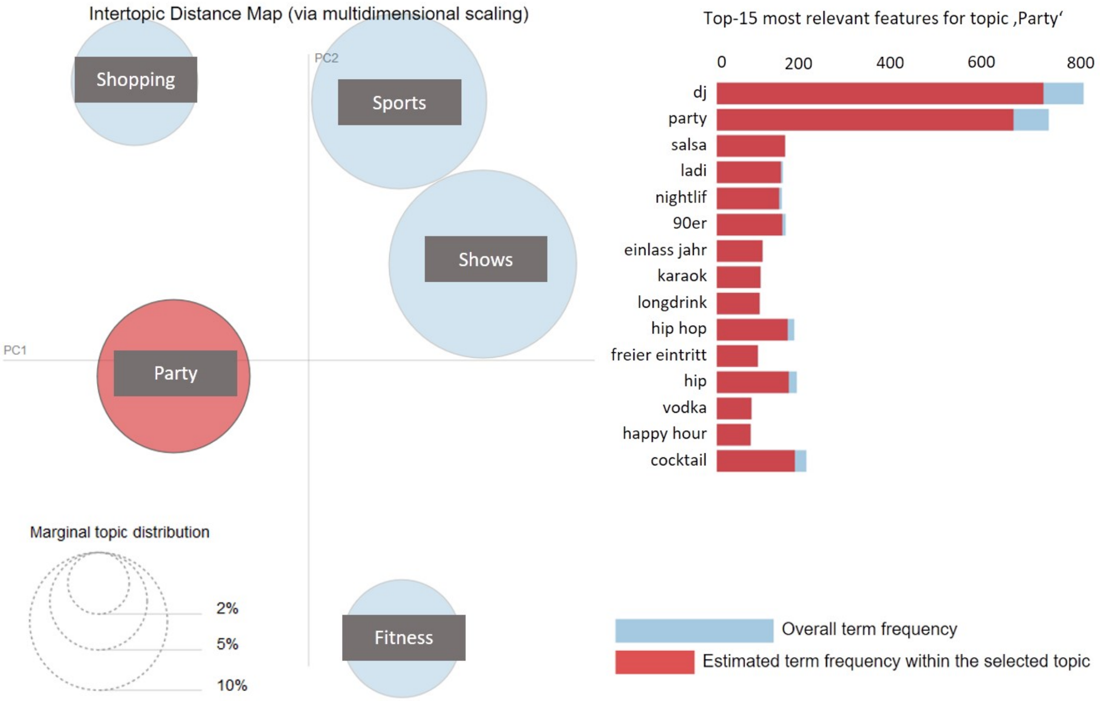


Figure 37. Topic distribution and most relevant features

Topic implications on modal split In order to highlight the relationship between parking events and the identified topics, textual data for Facebook events in the analysis area (Chapter 6.4.1) are transformed into tf-idf features. This involves about 27,000 event objects while all relevant text related to one object is treated as a separate document. A vectorizer module with 4,000 considered features is used. It is trained on the entire dataset to account for representative term relationships. Using the pre-trained LDA model with five topics, individual topic probabilities are assigned to each event object and the resulting vector is added to analysis dataset. Additionally, the number of parking events within the respective square polygons is added while parking events before, during and after the event are distinguished. The total number of parking events in the respective timeframes is subsequently separated into single features in accordance with the topic probabilities. For example, a probability of 30% for one of the topics is expected to indicate, that this topic is responsible for 30% of the observed parking events. Figure 38 shows the mean number of parking events within 0.5 km focus polygons organized by event topic and time. The analysis is conducted for both, 0.5 km and 2.0 km edge length of the focus polygons while identical patterns are observed. The results using 2.0 km polygons are found in Appendix F6.

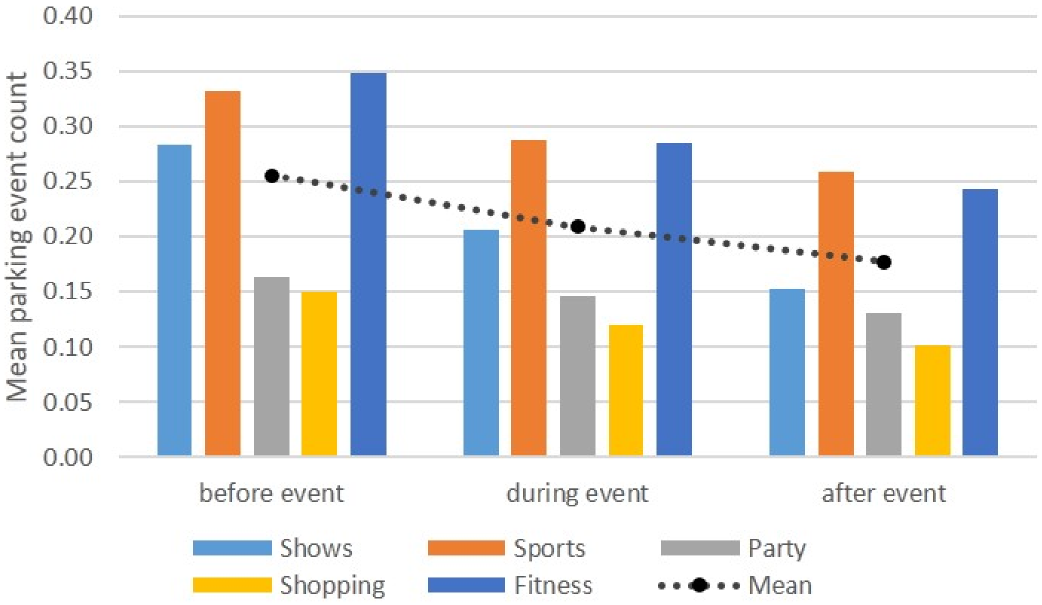


Figure 38. Probabilistic parking event distribution per topic (0.5 km focus polygons)

It can be seen that the topics ’Fitness’ and ’Sports’ generally involve a higher number of parking events than other topics. This finding is independent from the observed timewise event phase. 'Party' and 'Shopping' events involve comparably lower parking demand levels. The observed pattern for the topic 'Shows' is varied with regards to the different event timeframes. Respectively, 60 minutes before and after the respective start and end time of the event are considered. As indicated by the mean values over all topics, the number of parking events in the pre-event phase is generally higher than during the proceeding phases. While start times represent a mandatory field for event objects, most do not contain information for the end time attribute. Thus, an event duration of one hour is inserted for objects with missing end time values. As this estimation naturally deviates from the real end time, lower parking event counts within the tolerated 60-minute-timeframe are observed.

Moreover, it is assumed that many events also have flexible start and end times, have no scheduled activities or a strict attendance policy. On the one hand, attendees of trade fairs, clubbing events or shopping specials, for instance, can typically join the respective event at arbitrary times during the open timeframe. In this case, the start and end time supplied determines the maximum possible attendance time. On the other hand, events often require simultaneous group activity. For example, workshops or scheduled fitness trainings generally lead to clearer parking event patterns. The feature extraction workflow is applied to all event objects in the dataset. Topic probabilities are subsequently used as input for the main parking demand model.

6.5 Summary of extracted features

Figure 39 shows a summary of the features which were extracted using different methodologies described in this chapter. The respective number of attributes being added to the database is indicated for POIs (left) and events (right) for each technique applied.

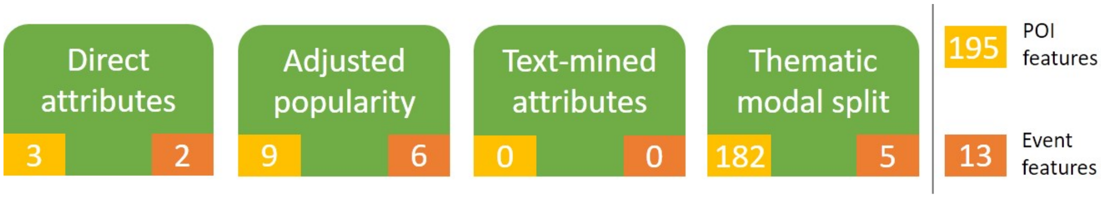


Figure 39. Summary of extracted features

Only some of the directly available attributes can be used for both categories. The adjusted popularity approach uses a reference data source for POIs and domain knowledge for events to increase the representativeness of the available Facebook data. Direct mining of text-based attributes is found infeasible due to a lack of textual data (POIs) or insufficient accuracy of preparatory classifiers (events). The extraction of features based on thematic object similarity is conducted using the specifically supplied category attribute on the POI side. This information is matched with historical parking events identified in FCD based on the 70 largest cities in Germany as an analysis area. As features, category-specific parking demand indicators are derived for both, based on POI opening hours and as specific values for combinations of weekdays and hours of the day. Conducting a similar approach is not possible for event objects as the majority of relevant category data is missing. Multi-class labeling models are found to achieve insufficient performance for being implemented as a corrective step. Thus, probabilistic category-assignment of event objects is conducted using LDA models. Here, the best model is selected based on the perplexity measure. It differentiates between five topics. The individual topic probabilities are computed for the analysis dataset, leading to five new input features being added. On the POI side, this procedure cannot be repeated due to the lack of textual information.

CHAPTER 7

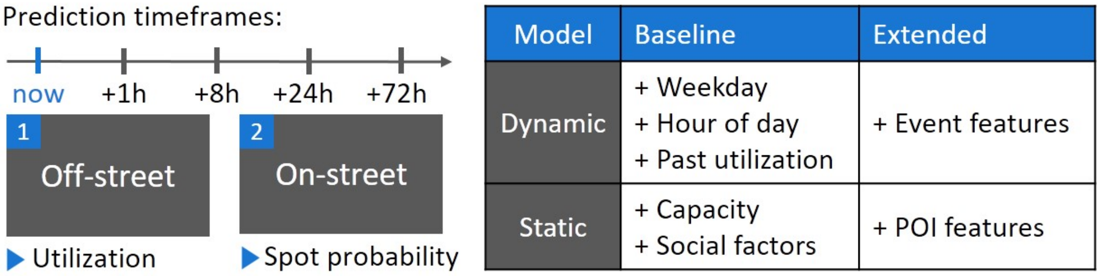
Feature evaluation

This chapter describes the procedures used for evaluation of the extracted input feature sets based on historical off- and on-street parking occupancy data. Multiple machine learning models are generated based on different feature constellations. A set of baseline features is introduced and the added value of the social media features is assessed.

7.1 Feature evaluation workflow

The first part of the evaluation is based on historical occupancy information related to diverse off-street facilities in multiple German cities. The target for this stage is high prediction performance on the relative utilization of the paid facilities. In the second stage, on-street parking utilization from long-term camera surveillance within the Braunschweig city area is used as ground truth. The first target value in this case is also the parking area utilization. The second target is represented by the binary differentiation between the parking area states 'full' and 'available'. Different timeframes for the parking occupancy predictions are applied. These comprise four levels ranging from short-term forecasts with one hour relative difference to long-term predictions being limited by a 72 hour timeframe. For each combination of application and prediction timeframe, separate models are trained including the social media feature sets to be evaluated. The obtained cross-validated model performances are subsequently compared to a baseline reference that considers only basic features. Figure 40 summarizes the described design of the feature evaluation workflow.

As dynamic features, the respective weekday, the hour of the day and past utilization are considered in the baseline model. The latter comprises a feature



Model	Baseline	Extended
Dynamic	+ Weekday + Hour of day + Past utilization	+ Event features
Static	+ Capacity + Social factors	+ POI features

Figure 40. Evaluation workflow and considered features

vector with the relative facility utilization five, 20 and 40 minutes prior to the respective data point. The extended model relies on added event features. In terms of static features, the respective parking area capacity, as well as diverse socioeconomic factors are considered for the baseline model. The latter comprises a set of merged data from the German Federal Statistical Office [169]. The following attributes are covered: Citizens per square kilometer, income per capita, cars per one thousand citizens, unemployment rate and gross domestic product per citizen. Furthermore, the respective share of the city area on the total area of the region as an expression of urbanization intensity, as well as demographic data, explicitly the share of people within the age range from 45 to 65 and over 65, are considered.

In the first phase, a baseline and an extended model are generated using only dynamic features. Here, all factors defining the context of the facility are implicitly included in the model. Due to their static character, all possible combinations of the dynamic factors happen under identical static influences. Thus, they add no value to facility-specific models. In the second phase, a generalized model over all facilities is trained using the static parameters as decisive elements. This bears the advantage of potential cross-learning among facilities and require less data per location.

As the extracted features are all object-specific, an aggregation procedure is necessary that summarizes the effects of multiple objects on the occupancy of parking areas in the surrounding. This is realized by computing the respective features over all objects in a focus polygon around the considered parking area. As the interaction of objects with regards to the target variables has not been sufficiently researched, potential effects in this area are neglected. In terms of polygon size, an edge length of 0.5 km is chosen due to the initial assumptions with regards to the potential walking distance of car drivers after having parked their car.

7.2 Off-street evaluation

In total, historical occupancy is available for 57 off-street facilities during the period from May 2016 to March 2017. As the memory requirements for the large extracted feature space represent a limitation of computability, a subset of six facilities is randomly selected for closer analysis.

In the first phase, different modeling approaches are compared for the generation of the baseline model. The set of tested algorithms includes SVMs with linear and radial basis kernel, an ANN regressor with sigmoid activation function and a random forest with 60 single estimators. The latter is found to achieve the best performance with 0.97 cross-validated R^2 for utilization predictions on the one-hour-timeframe and a 0.82 mean R^2 over all timeframes. Subsequently, the event features are added and the model is retrained. This leads to slight accuracy improvements with regards to the achieved mean R^2 of 0.84. The detailed results are provided in Figure 41.

It can be seen, that the prediction timeframe of one hour leads to high accuracies among all tested models. Longer prediction timeframes tend to result in lower model performances. However, better accuracies on the 24-hour-timeframe than for the eight-hour-timeframe are observed for all tested models. Adding event features is found to improve model performance by a small degree in both cases,

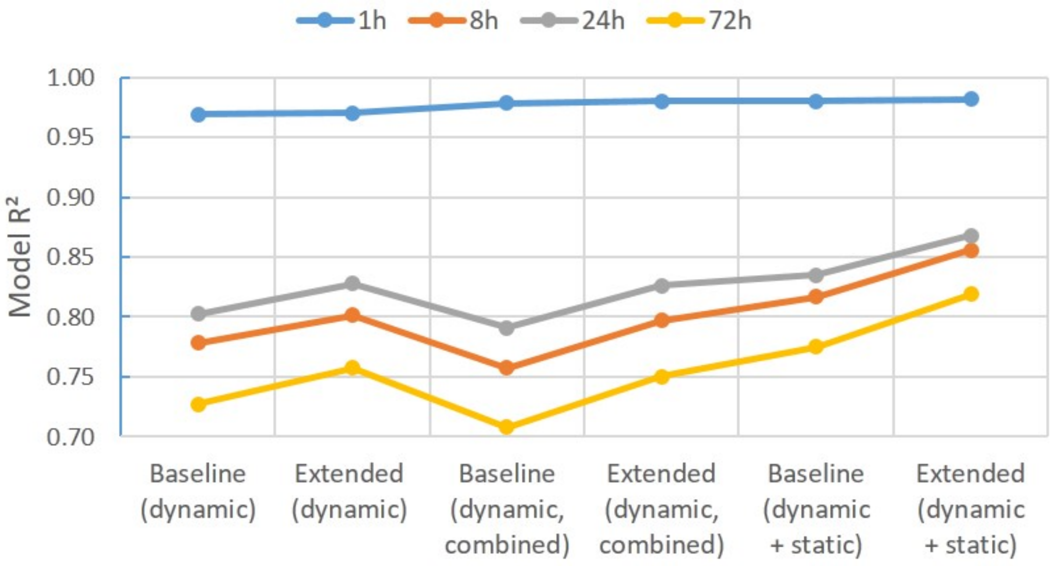


Figure 41. Random forest off-street prediction performance in different feature configurations

for the facility-specific models, as well as for the combined models using data over all tested facilities. As can be seen with regards to the analyzed model pair, the respective difference between baseline and extended model is larger for the combined model than for the mean over the single models. Regarding feature importances, the baseline features, in particular the past utilization attributes, are found to be most relevant. Considering event features, all attributes only have slight implications while the topic probabilities for currently active Facebook events are most important within this group.

Further accuracy improvements are achieved when the static feature sets are added to generate a combined model over all off-street locations. A mean R^2 of 0.85 is observed for the baseline model while the extended counterpart achieves a mean R^2 of 0.88. Predictions on the one-hour-timeframe achieve a MAPE of only 1.4% while 72-hour-predictions are conducted with 7.2% MAPE. Table 2 visualizes the relative variance explained by the single feature groups. Past utilization values are found to determine short-term occupancy predictions on the one-hour-timeframe while weekday and hour of day are more important for predictions eight hours in advance. In average over all timeframes, these two baseline groups combined are responsible for 86% of the observed variance. The importance of other features slightly increases for longer prediction periods but remains on a low level. To sum up, most POI and event features are found to be equally relevant and provide a slight benefit regarding model accuracy. However, the category-specific number of parking events within the opening hours of nearby places is not relevant. This also applies for the number of objects in the area as information resulting from aggregation and the facility-specific parking capacity.

Table 2. Explained variance by feature group and prediction timeframe for off-street occupancy

Feature group	1h	8h	24h	72h	Mean
Past utilization	96%	17%	59%	45%	54%
Weekday + Hour of day	2%	61%	29%	36%	32%
Social factors	0%	4%	2%	7%	3%
Hourly parking events	1%	8%	3%	5%	4%
Event topics	0%	2%	2%	3%	2%
Adjusted popularity	0%	4%	2%	3%	2%
Direct popularity	0%	2%	1%	2%	1%
Aggregation	0%	1%	0%	1%	0%
Opening hours parking events	0%	1%	0%	0%	0%

As the dimensionality of the extended feature set is extremely high, the computational resources for model training need to be considered. This effect is reduced by applying PCA to the dataset to reduce the dimensionality of the feature space and to create more valuable parameters. The observed model performances using the full feature space and using the PCA feature set with 15 constructed parameters are in an equal range. Univariate selection of features as a preparatory step to model generation is not necessary as the chosen random forest approach supports embedded selection.

7.3 On-street evaluation

Camera-based occupancy data is available for seven on-street parking areas within the Braunschweig city area. In total, about 36,500 data points are accessible in a five-minute-resolution. The availability of data ranges form about 3,500 to 7,000 occupancy points for a single parking area while the collection timeframe covers March and April 2017. The data is generated by capturing photos of the respective parking areas using battery-powered outdoor cameras with infrared-based night vision. Automated image analysis is used to extract the number of vehicles being located on the displayed parking area and the historical occupancy is calculated.

As data availability is critical, it is not possible to generate separate models for each parking area. This would lead to low generalization of the underlying feature relationships. Thus, only combined models over all locations are trained for the baseline and extended features. Here, the previously considered set of social factors is not taken into account as the underlying information is only available on a citywide basis. As all camera locations are within the same city, these features do no contribute any variance. Furthermore, street parking is usually not equipped with continuous acquisition of occupancy data. Thus, past utilization values, that facilitate the modeling of short-term patterns, are not equally available as for off-street. Thus, the respective feature set is excluded from the input data.

Figure 42 visualizes the observed performances of random forest models with 60 single estimators when using different feature sets and formulations of the target vector. When the usable dynamic and static baseline features are included into one combined model over all locations, a mean R^2 of 0.26 is observed. In this scenario, the on-street utilization is taken into account as a relative value between the numbers zero and one. Using the extended feature set with POI and event information, the model performance is increased to a 0.46 mean R^2 achieved. This is considered insufficiently accurate to be used in a productive information system that covers parking availability.

However, when searching for parking, it is actually only important for drivers to be able to distinguish between completely occupied and at least partly available parking areas. Thus, the granular occupancy information is translated into a binary feature vector that accounts for this information. Training models with binary target vector leads to significantly higher achieved accuracies. The baseline model shows a mean R^2 of 0.80 while the extended model shows further improvements up to a mean R^2 of 0.83.

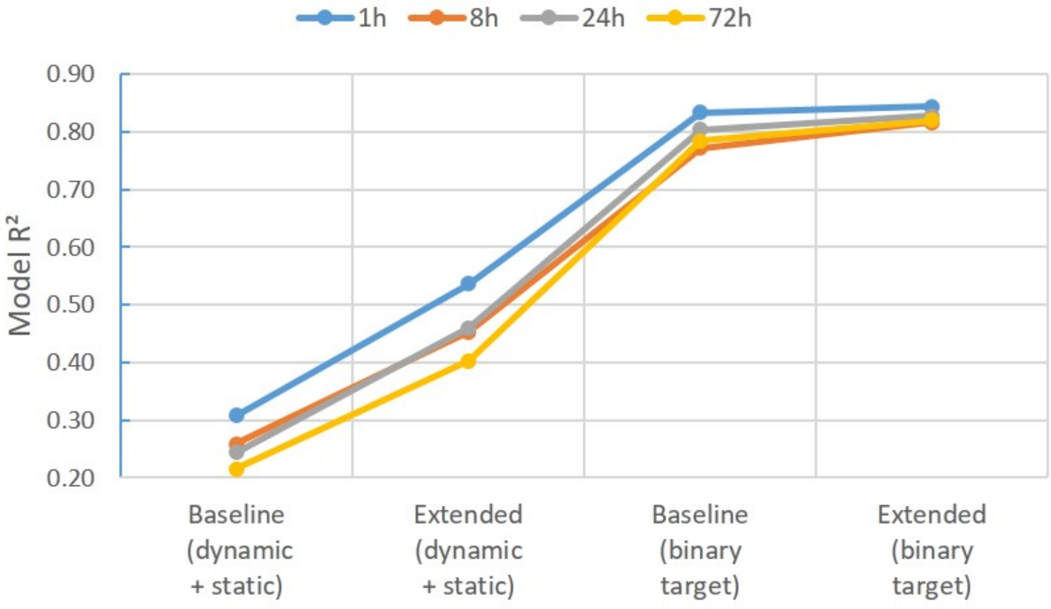


Figure 42. Random forest on-street prediction performance in different feature and target configurations

As parking-related traffic is a dynamic process, certain spots can be unavailable within an on-street setting even if no other car parks there. Vehicles can potentially block multiple spots or simply remain unrecognized by the camera system. Also, there is a certain changeover time to be considered for recently emptied spots to be reoccupied. To account for these circumstances, a flexible utilization threshold is tested that determines the binary parking availability. For example, when considering parking areas to be full over utilizations of 90% and more, a mean R^2 of 0.84 is achieved. Moreover, PCA is applied to the feature set, leading to slight improvements of the mean model R^2 up to 0.85. These results are based on 20 constructed PCA features.

Table 3 shows the respective feature importances for the best prediction model using the unchanged extended feature set. It can be seen that the weekday and hour of the day attributes determine most of the observed variance. The event topic features and the hourly parking events for POI focus areas are also found to be beneficial. The remaining feature sets are rather unimportant.

Table 3. Explained variance by feature group and prediction timeframe for on-street occupancy

Feature group	1h	8h	24h	72h	Mean
Weekday + Hour of day	35%	54%	50%	56%	49%
Event topics	53%	22%	26%	21%	31%
Parking events hourly	9%	18%	16%	14%	14%
Adjusted popularity	1%	1%	2%	3%	2%
Aggregation	1%	1%	2%	3%	2%
Parking events opening hours	1%	4%	2%	2%	2%
Direct popularity	0%	0%	0%	0%	0%

CHAPTER 8

Discussion

In the first phase, the research activities comprise the development of scalable methodologies for data acquisition and the exploratory analysis of the available data. Using a self-developed benchmarking methodology, the scope and quality of the data from social media is evaluated against alternative data sources. In the second phase, diverse information retrieval techniques are applied to extract potentially relevant sets of input features for the ultimate goal of parking demand modeling. Finally, the value-added by these feature is evaluated against a ground truth of parking occupancy data for both off- and on-street parking facilities.

8.1 Data acquistion

In the course of this thesis, leading online data sources in the areas of social media, mapping and events are preliminarily benchmarked based on publicly accessible popularity indicators. Facebook, OSM, Eventbrite and Ticketmaster are chosen as target platforms and scalable approaches are developed for data acquisition from these sources. While the collection of data based on publicly accessible web APIs is feasible at scale for most of the considered sources, data acquisition using the Facebook API requires the development of a specific methodology. It is based on achieving complete aerial coverage using a large number of API requests that supply location-specific data for a circular area. A flexible, density-based parametrization is implemented that leads to 1.41 million POI objects and 1.7 million event objects retrieved. The aerial size of zip code areas is used as the basis for estimating the location-specific degree of urbanization. Only urban areas are selected for data acquisition that represent 60% of the total aerial extent of Germany. The collected objects include a variety of metadata such as the object-specific online popularity and further attributes.

While it is likely that zip code areas do not indicate the respective object density in a highly accurate manner, no detailed demographic or alternative indicators are available that may serve as a ground truth or as a better indicator. Based on the limited availability of alternatives, it is reasonable to use this data as basis for the density estimation.

Generally, the developed collection algorithm is based on a large number of API calls that cause load on the provider side. It represents a workaround that specifically targets the proprietary structure of the API, allowing it to supply geographically-referenced data with sufficient coverage. Load limitations prohibit the efficient parallelization of API calls by requiring waiting time after a certain data volume has been extracted. Thus, the time consumption by the data acquisition process has to be taken into account when implementing it into a productive system. The observed changes over time regarding the extent of provided data are considered to be manageable with moderate cycles for data recollection. Thus, the acquisition procedure is considered sufficiently scalable.

8.2 Database benchmark

In the second phase, acquired reference data is compared to the Facebook dataset using a multi-stage procedure for identification of duplicate objects. This includes contextual matching with focus on geographic and time-based proximity, syntactic matching of object names and thematic matching of objects based on their respective category. For the name matching phase, a combined methodology is developed that is based on the longest common substring method and the Jaccard index as key similarity indicators. Categorical matching is based on manual assignment of congruent category labels among the data sources. This process is supported by syntactic label matching and dictionary-based retrieval of similarities. As none of the constructed indicators can determine alone whether two objects from different data sources are actually duplicates, they are all used as a combined feature vector with multiple, source-specific learning machines. Manually labeled data is generated and used as a basis for classifier training and testing. This leads to cross-validated duplicate identification accuracies of 97% on the POI side and a 19% identified overlap between the OSM and the Facebook dataset, representing 267,000 POI duplicates. As the extent of both sources is very large and the observed share of duplicates is comparably low, both data sources are considered valuable while the focus of this thesis lies on the social media side.

Regarding the overlap between Ticketmaster and Facebook events, an accuracy of 77% is achieved for the trained model, leading to 550 identified duplicates. Taking into account the overlap between Eventbrite and Facebook events, a classifier accuracy of 95% is observed, representing 635 duplicates. As the Facebook dataset for the same timeframe is multiple times larger than the benchmark sources and duplicates represent a large share of them, Facebook is considered to be fully superior. The availability of popularity data and textual object descriptions also supports this evaluation.

8.3 Feature engineering

8.3.1 Adjusted popularity measures

The literature review and exploratory data analysis indicates that the Facebook dataset is a skewed representation of the actual behavior observed in society. Thus, in terms of feature extraction, directly available popularity attributes are adjusted. One approach represents adjustment based on a reference data source that is assumed to be representatively distributed in terms of themes covered and user interaction observed. While OSM is used for this purpose on the POI side, no adequate reference source is available on the event side. The second adjustment-focused approach is based on the inclusion of domain knowledge from the literature, being applicable only on the event side due to the availability of relevant findings. As one of few published sources of information in this area, publications by the Japanese event platform Doorkeeper are taken considered for feature extraction. As it covers mostly professionally-themed and paid events within a different cultural context than the target area Germany, the representativeness of the information delivered remains unclear. However, these findings represent the most detailed source of information available as only very limited research has focused on this particular area.

8.3.2 Text mining for feature extraction

Another approach for feature extraction is developed that focuses on the explicit retrieval of attributes from textual contents of the Facebook objects. As the collected POIs do not contain a sufficient amount of text, only event objects are taken into account for this process. Text is transformed into term-based features using the tf-idf technique and modeling is conducted using multi-label machine learning classifiers. The target attributes to be extracted represent certain concepts indicated by the respective object text. One set of target labels focuses on demographics and behavioral characteristics of the event's potential attendees. Certain influence factors are covered that are presented to be relevant in literature for the travel mode choices made, implicitly indicating the parking demand for certain event objects. In particular, attributes denoting events specifically for elderly people, individuals with relatively high income and environmentally-aware users are introduced. However, low penetration of the dataset regarding events that fall in one of these categories makes these attributes impossible to be used in an automated classification context.

For this reason, alternative target attributes describing the actual event contents are added. The focus is set on events that are held outdoors and events that involve alcohol. Outdoor events are expected to imply a weather-dependent mobility behavior while alcohol is expected to shift the observed modal split away from car usage, decreasing the parking demand. Given a more prominent penetration of the dataset with these attributes, derived machine learning classifiers only achieve cross-validated accuracies up to 60%. As the error introduced by feature extraction models directly influences the subsequent parking demand modeling, the text-mining-based attribute retrieval is considered to be infeasible with the available scope of manually labeled data. An estimation of the required number of labeled samples is conducted that may lead to higher classifier accuracy. The approach is based on fitting a learning curve over models being trained on a variety of different sample sizes. It is found that adding further samples cannot meaningfully increase the classifier accuracies achieved. This behavior is interpreted as a result of strong data heterogeneity regarding the number of common term features in different objects. Adding new samples mainly leads to diverging content being added and a higher number of distinct words being taken into account. Compared to popular reference corpora in the natural language processing literature, the event-related data is in fact observed to be more heterogeneous and less easily generalizable.

In this phase, a fixed set of stop words and stemming as preparation of terms are used for all classifications. Lemmatization-based feature preparation was also tested and found to increase the heterogeneity of the training-ready feature set. This is explained by the characteristic differentiation of semantically similar terms based on their diverging suffixes. In fact, this alternative preparation technique is less favorable. Also, alternatives to tf-idf features like the binary consideration of term occurrences are tested but found to lead to lower classifier accuracies.

8.3.3 Thematic modal split modeling

Another feature extraction approach developed focuses on creating indicators for parking demand based on the thematic character of the analyzed Facebook objects. These parking events from FCD are considered as direct modal split indicator with particular focus on car usage. Focus areas in the 70 largest cities of Germany are defined and used to draw connections between aggregated parking events and the Facebook objects. On the POI side, parking events during and after opening hours are aggregated for smaller focus areas around each object to create category-specific features. This data gives clear indications on traffic and particularly parking demand in the respective time frames and allows partly isolating the effect of POIs in the focus areas by comparing open and closed timeframes.

However, the approach introduces bias based on the observation that opening hours for many POI categories primarily cover active daytime with an independently greater traffic demand compared to the nighttime hours. In fact, the parking-related influence of POIs cannot be fully separated from other factors such as the weekday- and hour-of-day-specific background influence. Moreover, 70% of all objects are missing opening hour information and this type of feature cannot be extracted, which strongly limits the amount of data for pattern recognition. Also, due to limitations in computability, only POI objects in the focus areas are considered. With over 1,100 different POI categories in the Facebook dataset, the number of objects considered in the focus areas is not large enough. Smaller object counts are assumed to limit the generalization of potentially underlying patterns. As a minimum of category-specific samples is required, only about 26% of all POI categories can be taken into account for feature extraction. This coverage is generally considered to be insufficient for meaningful feature extraction but the created set is still passed to the feature evaluation phase.

It is observed that POIs in the same category have very similar opening hours. Thus, instead of projecting the underlying patterns only on a binary system that distinguishes between open and closed POI objects, parking events are aggregated on an hourly basis for each day of the week. This allows further detail in explaining the observed variance related to specific object categories. Similar to the previous approach, the scope of covered categories is limited. Based on the level of detail applied in the Facebook category scheme, the mean number of historical parking events for a certain category per combination of weekday and hour of the day is extremely low and does not reflect meaningful patterns for the intermediate surrounding areas. Thus, object focus areas larger than the assumed walkable distance are constructed to create sufficient differentiation within the parking event feature set. It remains unclear whether parking events in the extended focus areas actually represent object-related patterns or are rather determined by external influences.

Furthermore, it has to be noted that the analyzed FCD is not evaluated for its capability to represent real-life traffic patterns with sufficient accuracy. The FCD is provided by external suppliers that potentially only cover user groups with certain behavioral patterns. It is possible that peaks are observed at locations where the source is more popular than at other points. As no reasonable statement can be made with regards to these aspects, an unbiased representativeness is assumed for feature generation in the given applications. Additionally, the presented processes do not distinguish between different kinds of parking events. Individual patterns are observed for vehicles leaving their parking spot, successful parking and unsuccessful searching for parking. These may have different implications on the availability of parking space. The main reason for neglecting this differentiation is a limited number of available parking event data in the respective categories if they are assigned to specific Facebook objects. Only the summarized version allows to distinguish object-specific patterns. Finally, only parking events from a timeframe of two months are put into consideration for computational reasons. It remains unclear whether the observed patterns in this period adequately represent potential general observations on yearly basis.

The category attribute is not available for most Facebook event objects. Thus, the feature extraction approaches on the POI side cannot be directly applied for events. First, a set of text-based predictors is developed that aims at replacing the missing categorical attribute. Different levels of thematic summarization are tested but no sufficiently accurate classifier could be constructed based on the generated set of labeled data. Thus, unsupervised topic modeling based on LDA is applied on the event-related text data. This technique identifies thematic concepts in the entire text corpora in a probabilistic manner. Continuously evaluating the degree of model generalization, a configuration distinguishing five topics is selected. Subsequently, each event in the focus dataset is assigned with respective topic probabilities and the observed number of parking events is analyzed in relation to the distinguished topics. Event objects with a high probability for the themes 'sports' and 'fitness' are found to correspond with higher parking intensities. The respective probability vector is directly used as feature input.

8.4 Feature evaluation

The extracted features are evaluated for their implications on the occupancy of parking space at selected locations in Germany. Various prediction timeframes are used that range from one to 72 hours in the future from the momentary situation. For each configuration, the performance of a baseline model using a set of basic features is compared to an extended model that has the extracted social media features added. Focus areas covering a walkable distance around the considered parking locations are defined and object-related features within these areas are summarized to reflect co-existence. Potentially important interaction patterns among objects of certain themes or categories are not investigated. In the literature, the mobility-related interaction of POI or event objects has hardly been discussed and simple feature summarization is applied based on the lack of more promising approaches. For each feature configuration, different learning algorithms are tested and random forest models are found to outperform the evaluated alternatives.

On the off-street side, a mean R^2 over all tested prediction timeframes of 0.88 is achieved based on the extended feature set including social media features. The corresponding baseline model only leads to a mean R^2 of 0.85. Short-term predictions for one hour in the future strongly rely on past utilization values from the past hour. In fact, there are local short-term trends that can be identified and modeled with this feature set. For longer prediction timeframes, also the respective weekday and hour of the day turn out to be particularly relevant. On average, over all prediction timeframes, these two sets remain important and explain about 86% of the observed variance. Direct and adjusted popularity measures, event topics and hourly parking events account for an average 9% of the observed variance only. Even though the efforts for feature extraction are high, valuable accuracy improvements are induced by certain social media feature groups.

Regarding on-street parking areas, camera-based occupancy data from strategically important locations in the Braunschweig city area is used for model training. The extended feature set is similarly used to forecast the share of spots utilized within the monitored parking areas. A mean R^2 of 0.46 is achieved, representing a significant increase from the mean R^2 of 0.26 for the baseline model. However, both models are not sufficiently accurate to derive valuable occupancy predictions.

Thus, the problem formulation is transformed into modeling the binary state of parking areas 'full' and 'available'. This is expected to reflect the actual information need of drivers in a more user-centered way. For them, it is only relevant whether or not parking spots are available at a specific location and time. Based on the binary target classes, random forest classifiers are trained with the previously used feature sets. The baseline model leads to a mean R^2 of 0.80 while the extended model triggers improvements up to a 0.83 mean R^2. Dimensionality reduction of the feature space with PCA leads further accuracy increases up to a 0.85 mean model R^2. While weekday and hour of day as input features are still of superior importance, the event topic features are responsible for 31% of the explained variance. The hourly parking event features account for 14% of the explained variance. Other feature sets show no significant contribution. Past utilization values are excluded as these near real-time data points are typically not available in a productive forecasting system. This kind of information can only be supplied by on-street parking sensors that are only available for very limited urban areas.

It has to be noted that the on-street findings are based on a rather small ground truth as no extensive occupancy data for these areas is available. This increases the risk of observed model overfitting as no broad set of parameter configurations can be used for training and testing. Also, all observation points are located within the Braunschweig city area, preventing statistical social factors to be included. It is possible that validating an on-street occupancy model with this data leads to geographic overfitting on the tested city. This is the case if traffic and car usage patterns in other cities are fundamentally different. However, based on the homogeneity of parking-related findings in the literature over different geographical contexts, similarity among cities can be assumed.

8.5 Conclusion and contributions

This thesis represents one of the first studies that focuses on large-scale feature extraction from social media to model urban parking demand. It draws one of the first connections between crowdsourced data and modal split using extracted data from FCD and various other sources with extensive coverage. Multiple approaches for scalable data acquisition and an accurate methodology for text-based identification of duplicates in heterogeneous online databases are introduced. Here, an extension of established procedures for syntactic similarity mining is developed and applied for benchmarking of social media against alternative online data sources. Findings in the literature are used to identify potentially relevant modeling parameters that are covered with specifically extracted attributes from the raw data. Among others, this phase covers the adjustment of directly available popularity attributes based on reference data sources and external publications. Also, text- and model-based identification of the targeted attendee group and further event characteristics are tested. This approach was found to be infeasible with comparably small labeled datasets due to the heterogeneity and divergence of the event text corpus. Finally, thematic similarities among POIs and events are used to draw category-based connections to historical parking events extracted from FCD. An extensive analysis area covering the 70 largest cities in Germany is used to derive thematic features. For events, the text-based reconstruction of missing attributes is tested and finally, an unsupervised topic modeling based on LDA is applied to derive probabilistic features focusing on thematic similarity.

The evaluation of constructed features is based on historical data for multiple off-street facilities across Germany and an on-street ground truth for the city of Braunschweig. Separate models were generated using a baseline selection of influence features, as well as using an extended feature set comprising extracted social media attributes. Random forest models were found to perform best among different tested learning algorithms, leading to a mean R^2 of 0.88 over different prediction timeframes for off-street facilities with the extended feature set. Here, the extracted social media features were found to explain a low, but still relevant part of the observed variance. For the tested on-street facilities, a mean model R^2 of 0.85 is achieved using PCA for feature preparation and a binary target attribute that distinguishes between available and not available parking areas. Here, event topic probabilities and aggregated parking events on an hourly basis are identified as particularly relevant input sets. Summarizing, it is recommended to include social media features into parking demand modeling as their integration leads to comparably small, but valuable accuracy improvements of the underlying machine learning models.

8.6 Future work

For the future, it is recommended to extract and test further feature configurations from the raw social media data. The integration of further data sources and potential future findings from the literature may lead to further accuracy improvements. For instance, as benchmarking showed that OSM is a comprehensive data source showing comparably low duplicates compared to the Facebook POI set, future research may focus on the integration of derived features using OSM as a data source. POI-related text data may be cross-referenced from secondary online sources to create the basis for text mining and popularity estimation.

New findings regarding the representativeness of social media in comparison to the physical attendance behavior in society may lead to improvements of the adjusted popularity approach. In this area, huge potential is seen with regards to more in-depth data covering the interaction behavior of social media users with POIs and events. The most promising approach is to build up a large focus group that voluntarily contributes information for behavioral analysis based on social media data. Besides opening up interesting research potential in the social sciences, these behavioral patterns are expected to be a valuable basis to understand how interactions in social media have implications on the individual mobility behavior in real life. This may include many parking-related influences such as age, income, car accessibility and particular interests. Exemplarily, this may allow to gasp the social sentiments towards particular events in order to derive more accurate estimates for the observed offline attendance. It is possible to include individual behavior as part of an agent-based simulation or a similar technique. Also, social media mining may be extended by analyzing photo and video data to recognize implicitly happening events and get estimates on their mobility implications. Having access to highly detailed data of individuals, their media usage and travel mode choices, provides the opportunity to derive varied new features for parking demand modeling. Closing the research gap between highly available social media data and individual mobility behavior is expected to have a significant impact on many areas such as city planning and digitalized mobility services.

Furthermore, a larger on-street ground truth for parking occupancy covering more data points and more diverse locations would provide a generally higher reliability of the derived findings. As the generalization of on-street occupancy models is directly dependent on this data, representativeness is particularly important. Moreover, larger sample sizes for the text-based attribute extractors may lead to higher achieved accuracy of the generated models. This requires significant further labeling and data acquisition over longer timeframes. Given the required computational resources, larger amounts of FCD can be taken into account to derive revised feature sets. In this case, the applied focus areas for deriving features may also be extended to a nationwide scale. A larger dataset for topic modeling may also lead to different, potentially more valuable thematic structures being identified in the text corpora.

Finally, evaluating the representativeness of the available FCD sources in comparison to other traffic intensity indicators may increase the degree of data understanding and reliability of derived models. Also, further applications of parking events and other comparable features from FCD can be researched.

LIST OF REFERENCES

[1] Ministry of Public Security China and Xinhua News Agency, "Vehicle population in china from 2007 to 2015," 2016. [Online]. Available: https://www.statista.com/statistics/285306/number-of-car-owners-in-china/

[2] "Parken 2020: Quantum fokus." [Online]. Available: https://www.quantum.ag/fileadmin/Dateien/Publikationen_Archiv/QuantumFokus_2-2012.pdf

[3] X. Chen and N. Liu, "Smart parking by mobile crowdsensing," *International Journal of Smart Home*, vol. 10, no. 2, pp. 219–234, 2016.

[4] "Smart parking: A system that could help cities rethink parking: Background information," 2015. [Online]. Available: http://www.siemens.com/press/pool/de/events/2015/corporate/2015-09-iaa/background-smart-parking-e.pdf

[5] C. Morillo and J. M. Campos, "On-street illegal parking costs in urban areas," *Procedia - Social and Behavioral Sciences*, vol. 160, pp. 342–351, 2014.

[6] D. C. Shoup, "Cruising for parking," *Transport Policy*, vol. 13, no. 6, pp. 479–486, 2006.

[7] S. Belloche, "On-street parking search time modelling and validation with survey-based data," *Transportation Research Procedia*, vol. 6, pp. 313–324, 2015.

[8] D. Pfoser, "Floating car data," in *Encyclopedia of GIS*, S. Shekhar and H. Xiong, Eds. Boston, MA: Springer US, 2008, p. 321.

[9] M. Caliskan, D. Graupner, and M. Mauve, "Decentralized discovery of free parking places," in *Proceedings of the 3rd international workshop on Vehicular ad hoc networks*, W. Holfelder and D. Johnson, Eds., 2006, p. 30.

[10] G. Surpris, D. Liu, and D. Vincenzi, "How much can a smart parking system save you?" *Ergonomics in Design: The Quarterly of Human Factors Applications*, vol. 22, no. 4, pp. 15–20, 2014.

[11] F. Caicedo, F. Robuste, and A. Lopez-Pita, "Parking management and modeling of car park patron behavior in underground facilities," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1956, pp. 60–67, 2006.

[12] A. Sakai, K. Mizuno, T. Sugimoto, and T. Okuda, "Parking guidance and information systems," in *Pacific Rim TransTech Conference. 1995 Vehicle Navigation and Information Systems Conference Proceedings. 6th International VNIS. A Ride into the Future*, 1995, pp. 478–485.

[13] S. A. El-Seoud, H. El-Sofany, and I. Taj-Eddine, "Towards the development of smart parking system using mobile and web technologies," in *2016 International Conference on Interactive Mobile Communication, Technologies and Learning (IMCL)*, 2016, pp. 10–16.

[14] S. Brooke, S. Ison, and M. Quddus, "On-street parking search," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2469, pp. 65–75, 2014.

[15] "Knowing where to park without looking," 2016. [Online]. Available: http://www.siemens.com/press/en/events/2015/mobility/2015-09-smart-parking.php?content[]=Corp&content[]=MO

[16] M. Tschentscher, C. Koch, M. Konig, J. Salmen, and M. Schlipsing, "Scalable real-time parking lot classification: An evaluation of image features and supervised learning algorithms," in *2015 International Joint Conference on Neural Networks (IJCNN)*, 2015, pp. 1–8.

[17] J. Wolff, T. Heuer, H. Gao, M. Weinmann, S. Voit, and U. Hartmann, "Parking monitor system based on magnetic field sensors," in *2006 IEEE Intelligent Transportation Systems Conference*, 2006, pp. 1275–1279.

[18] S. Mathur, T. Jin, N. Kasturirangan, J. Chandrasekaran, W. Xue, M. Gruteser, and W. Trappe, "Parknet: drive-by sensing of road-side parking statistics," in *Proceedings of the 8th international conference on Mobile systems, applications, and services*, S. Banerjee, S. Keshav, and A. Wolman, Eds., 2010, p. 123. [Online]. Available: 10.1145/1814433.1814448

[19] M. Rinne and S. Törmä, "Mobile crowdsensing of parking space using geofencing and activity recognition," in *2014 10th European ITS conference*, 2014. [Online]. Available: http://s3.amazonaws.com/academia.edu.documents/40832996/Mobile_crowdsensing_of_parking_space_using_geofencing_and_activity_recognition_final.pdf?AWSAccessKeyId=AKIAJ56TQJRTWSMTNPEA&Expires=1478549535&Signature=rmVTur6nzu46eZl%2BEHMqDk53EHY%3D&response-content-disposition=inline%3B%20filename%3DMobile_crowdsensing_of_parking_space_usi.pdf

[20] F. J. Villanueva, D. Villa, M. J. Santofimia, J. Barba, and J. C. Lopez, "Crowdsensing smart city parking monitoring," in *2015 IEEE 2nd World Forum on Internet of Things (WF-IoT)*, 2015, pp. 751–756.

[21] J. Riedel, "Dynamic parking space surveillance using magnetic field sensors of smartphones," Independent study, University of Rhode Island, Kingston, RI, USA, 2016.

[22] R. Hössinger, P. Widhalm, M. Ulm, K. Heimbuchner, E. Wolf, R. Apel, and T. Uhlmann, "Development of a real-time model of the occupancy of short-term parking zones," *International Journal of Intelligent Transportation Systems Research*, vol. 12, no. 2, pp. 37–47, 2014.

[23] R. Wirth and J. Hipp, "Crisp-dm: Towards a standard process model for data mining," in *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining*, 2000, pp. 29–39.

[24] "Crisp-dm 1.0: Step-by-step data mining guide." [Online]. Available: https://www.the-modeling-agency.com/crisp-dm.pdf

[25] G. Seni and J. F. Elder, "Ensemble methods in data mining: Improving accuracy through combining predictions," *Synthesis Lectures on Data Mining and Knowledge Discovery*, vol. 2, no. 1, pp. 1–126, 2010.

[26] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, no. 3, pp. 1157–1182, 2003. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=9082CF990A78224DCF6B16FC8BEF1A6A?doi=10.1.1.3.8934&rep=rep1&type=pdf

[27] J. Heaton, "An empirical analysis of feature engineering for predictive modeling," in *SoutheastCon 2016*. IEEE, 2016, pp. 1–6.

[28] H.-f. Yu, Lo Hung-yi, H.-p. Hsieh, J.-k. Lou, T. G. Mckenzie, J.-w. Chou, P.-h. Chung, C.-h. Ho, C.-f. Chang, J.-y. Weng, E.-s. Yan, C.-w. Chang, T.-t. Kuo, P. T. Chang, C. Po, C.-y. Wang, Y.-h. Huang, Y.-x. Ruan, Y.-s. Lin, S.-d. Lin, H.-t. Lin, and C.-j. Lin, "Feature engineering and classifier ensemble for kdd cup 2010," in *In JMLR Workshop and Conference Proceedings*, 2011.

[29] J. Brownlee, "Discover feature engineering, how to engineer features and how to get good at it," 2014. [Online]. Available: http://machinelearningmastery.com/discover-feature-engineering-how-to-engineer-features-and-how-to-get-good-at-it/

[30] M. Boullé, "Towards automatic feature construction for supervised classification," in *Machine Learning and Knowledge Discovery in Databases*, ser. Lecture Notes in Computer Science, T. Calders, F. Esposito, E. Hüllermeier, and R. Meo, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2014, vol. 8724, pp. 181–196.

[31] K. A. Ross, C. S. Jensen, R. Snodgrass, C. E. Dyreson, S. Skiadopoulos, C. Sirangelo, M. L. Larsgaard, G. Grahne, D. Kifer, H.-A. Jacobsen, H. Hinterberger, A. Deutsch, A. Nash, K. Wada, W. M. P. Aalst, C. Dyreson, P. Mitra, I. H. Witten, B. Liu, C. C. Aggarwal, M. T. Özsu, C. Ogbuji, C. Patel, C. Weng, A. Wright, A. Shabo, D. Russler, R. A. Rocha, Y. A. Lussier, J. L. Chen, M. J. Zaki, A. Corral, M. Vassilakopoulos, D. Gunopulos, D. Wolfram, S. Venkatasubramanian, M. Vazirgiannis, I. Davidson, S. Sarawagi, L. Peyton, G. Speegle, V. Vianu, D. van Gucht, O. Etzion, F. Curbera, A. Ericsson, M. Berndtsson, J. Mellin, P. M. D. Gray, G. Trajcevski, O. Wolfson, P. Scheuermann, C. Dorai, M. Weiner, A. Borgida, J. Mylopoulos, G. Vossen, A. Reuter, V. Tannen, S. Elnikety, A. Fekete, L. Bertossi, F. Geerts, W. Fan, T. Westerveld, C. Gurrin, J. Kekäläinen, P. Arvola, M. Junkkari, K. Mouratidis, J. X. Yu, Y. Yao, J. Gehrke, S. Babu, N. Palmer, C. K.-S. Leung, M. W. Carroll, A. Gokhale, M. Ouzzani, B. Medjahed, A. K. Elmagarmid, S. Manegold, G. Cormode, S. Mankovskii, D. Zhang, T. Härder, W. Gao, C. Niu, Q. Li, Y. Yang, P. Refaeilzadeh, L. Tang, H. Liu, T. B. Pedersen, K. Morfonios, Y. Ioannidis, M. H. Böhlen, R. T. Snodgrass, and L. Chen, "Curse of dimensionality," in *Encyclopedia of Database Systems*, L. LIU and M. T. Özsu, Eds. Boston, MA: Springer US, 2009, pp. 545–546.

[32] R. Rojas, "The curse of dimensionality," Berlin, 2015. [Online]. Available: https://www.inf.fu-berlin.de/inst/ag-ki/rojas_home/documents/tutorials/dimensionality.pdf

[33] L. van der Maaten, E. Postma, and H. van den Herik, "Dimensionality reduction: A comparative review," Ph.D. dissertation, Tilburg University, Tilburg, 2009. [Online]. Available: https://www.tilburguniversity.edu/upload/59afb3b8-21a5-4c78-8eb3-6510597382db_TR2009005.pdf

[34] A. R. S. A. Alamdari, "Variable selection using correlation and single variable classifier methods: Applications," in *Feature Extraction*, ser. Studies in Fuzziness and Soft Computing, I. Guyon, M. Nikravesh, S. Gunn, and L. A. Zadeh, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, vol. 207, pp. 343–358.

[35] J. Han, M. Kamber, and J. Pei, *Data mining: Concepts and techniques*, 3rd ed., ser. The Morgan Kaufmann series in data management systems. Amsterdam: Elsevier/Morgan Kaufmann, 2012. [Online]. Available: http://site.ebrary.com/lib/hamburg/Doc?id=10483440

[36] S. García, J. Luengo, and F. Herrera, *Data preprocessing in data mining*, ser. Intelligent Systems Reference Library. Cham: Springer, 2015, vol. 72. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-10247-4

[37] C. C. Aggarwal and C. Zhai, Eds., *Mining Text Data*, 2012nd ed. Boston, MA: Springer US, 2012. [Online]. Available: http://dx.doi.org/10.1007/978-1-4614-3223-4

[38] C. D. Manning, P. Raghavan, and H. Schutze, *Introduction to Information Retrieval.* Cambridge: Cambridge University Press, 2008.

[39] D. M. Blei, A. Y. Ng, M. I. Jordan, and J. Lafferty, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, p. 2003, 2003.

[40] S. Qian, T. Zhang, C. Xu, and J. Shao, "Multi-modal event topic model for social event analysis," *IEEE Transactions on Multimedia*, vol. 18, no. 2, pp. 233–246, 2016.

[41] "Die bereitstellung von daten über govdata," 2017. [Online]. Available: https://www.govdata.de/web/guest/datenbereitsteller

[42] H. Zhang and J. Malczewski, "Quality evaluation of volunteered geographic information," in *Volunteered Geographic Information and the Future of Geospatial Data*, ser. Advances in Geospatial Technologies, N. Dey, C. E. Calazans Campelo, M. Bertolotto, and P. Corcoran, Eds. IGI Global, 2017, pp. 19–46.

[43] C. Capineri, "The nature of volunteered geographic information," in *European Handbook of Crowdsourced Geographic Information*, C. Capineri, M. Haklay, H. Huang, V. Antoniou, J. Kettunen, F. Ostermann, and R. Purves, Eds. Ubiquity Press, 2016, pp. 15–33.

[44] O. Roick and S. Heuser, "Location based social networks - definition, current state of the art and research agenda," *Transactions in GIS*, vol. 15, no. 5, pp. n/a–n/a, 2013.

[45] M. Haklay, "Why is participation inequality important?" in *European Handbook of Crowdsourced Geographic Information*, C. Capineri, M. Haklay, H. Huang, V. Antoniou, J. Kettunen, F. Ostermann, and R. Purves, Eds. Ubiquity Press, 2016, pp. 35–44.

[46] N. Budhathoki, "Participants' motivations to contribute geographic information in an online community," Dissertation, University of Illinois, 2010. [Online]. Available: http://hdl.handle.net/2142/16956

[47] D. Jonietz and A. Zipf, "Defining fitness-for-use for crowdsourced points of interest (poi)," *ISPRS International Journal of Geo-Information*, vol. 5, no. 9, p. 149, 2016.

[48] C. Capineri, M. Haklay, H. Huang, V. Antoniou, J. Kettunen, F. Ostermann, and R. Purves, Eds., *European Handbook of Crowdsourced Geographic Information.* Ubiquity Press, 2016.

[49] M. Campagna, “Social media geographic information: Why social is special when it goes spatial?” in *European Handbook of Crowdsourced Geographic Information*, C. Capineri, M. Haklay, H. Huang, V. Antoniou, J. Kettunen, F. Ostermann, and R. Purves, Eds. Ubiquity Press, 2016, pp. 45–54.

[50] S. Kemp, “Digital in 2017: Global overview,” 2017. [Online]. Available: https://wearesocial.com/blog/2017/01/digital-in-2017-global-overview

[51] “Forecast of the social network user penetration rate in germany from 2014 to 2021,” 2017. [Online]. Available: https://www.statista.com/statistics/567322/predicted-social-network-user-penetration-rate-in-germany/

[52] “Forecast of facebook user numbers in germany from 2015 to 2021 (in million users),” 2016. [Online]. Available: https://www.statista.com/statistics/568790/forecast-of-facebook-user-numbers-in-germany/

[53] “Content marketing gains momentum in finland,” 2014. [Online]. Available: https://www.emarketer.com/Article/Content-Marketing-Gains-Momentum-Finland/1010912

[54] “Population: Germany, reference date, age: 12411-0005,” 2017. [Online]. Available: https://www-genesis.destatis.de

[55] “Number of facebook users in germany in january 2011 and 2014, by age group (in millions),” 2014. [Online]. Available: https://www.statista.com/statistics/451414/facebook-users-by-age-group-germany/

[56] W. Canzler, “Räumliche mobilität und regionale unterschiede: Auszug aus dem datenreport 2016.” [Online]. Available: https://www.destatis.de/DE/Publikationen/Datenreport/Downloads/Datenreport2016Kap11.pdf?___blob=publicationFile

[57] “Mobilität in deutschland 2008: Kurzbericht.” [Online]. Available: http://www.mobilitaet-in-deutschland.de/pdf/MiD2008_Kurzbericht_I.pdf

[58] “Anteil der nutzer von social media in deutschland nach bundesländern im jahr 2016,” 2017. [Online]. Available: https://de.statista.com/statistik/daten/studie/210763/umfrage/nutzung-von-social-media-in-deutschland-nach-bundeslaendern/

[59] “Distribution of facebook users worldwide as of january 2017, by age and gender,” 2017. [Online]. Available: www.statista.com/statistics/376128/facebook-global-user-age-distribution

[60] F. McAndrew and H. S. Jeong, “Who does what on facebook? age, sex, and relationship status as predictors of facebook use,” *Computers in Human Behavior*, vol. 28, no. 6, pp. 2359–2365, 2012.

[61] "Percentage of u.s. internet users who use facebook in april 2016, by annual household income," 2016. [Online]. Available: https://www.statista.com/statistics/246222/share-of-us-internet-users-who-use-facebook-by-household-income/

[62] R. E. Wilson, S. D. Gosling, and L. T. Graham, "A review of facebook research in the social sciences," *Perspectives on psychological science : a journal of the Association for Psychological Science*, vol. 7, no. 3, pp. 203–220, 2012.

[63] P. B. Brandtzaeg and I. M. Haugstveit, "Facebook likes: A study of liking practices for humanitarian causes," *International Journal of Web Based Communities*, vol. 10, no. 3, p. 258, 2014.

[64] "Reasons facebook users in germany have liked a company on facebook as of june 2012," 2012. [Online]. Available: https://www.statista.com/statistics/244907/reasons-facebook-users-in-germany-have-liked-a-company/

[65] "Reasons for becoming a brand fan on facebook," 2013. [Online]. Available: http://www.syncapse.com/wp-content/uploads/2013/06/Graph21.jpg

[66] S. Patil, G. Norcie, A. Kapadia, and A. J. Lee, "Reasons, rewards, regrets: Privacy considerations in location sharing as an interactive practice," in *Proceedings of the Eighth Symposium on Usable Privacy and Security - SOUPS '12*, L. F. Cranor, Ed. New York, New York, USA: ACM Press, 2012, p. 1.

[67] I. Bilogrevic, K. Huguenin, S. Mihaila, R. Shokri, and J.-P. Hubaux, "Predicting users' motivations behind location check-ins and utility implications of privacy protection mechanisms," in *Proceedings 2015 Network and Distributed System Security Symposium*, E. Kirda, Ed. Reston, VA: Internet Society, 2015.

[68] I. Polakis, S. Volanis, E. Athanasopoulos, and E. P. Markatos, "The man who was there: Validating check-ins in location-based services," in *Proceedings of the 29th Annual Computer Security Applications Conference on - ACSAC '13*, C. N. Payne, Ed. New York, New York, USA: ACM Press, 2013, pp. 19–28.

[69] "Verkehr und mobilität in deutschland: Daten und fakten kompakt." [Online]. Available: https://www.bmvi.de/SharedDocs/DE/Publikationen/G/verkehr-und-mobilitaet-in-deutschland.pdf?___blob=publicationFile

[70] "Gleitenden mittelfristprognose für den güter- und personenverkehr: Mittelfristprognose winter 2016/2017," Waldkirch / Köln. [Online]. Available: https://www.bag.bund.de/SharedDocs/Downloads/DE/Verkehrsprognose/Verkehrsprognose_winter_2016_2017.pdf

[71] "Daten zum verkehr." [Online]. Available: https://www.umweltbundesamt.de/sites/default/files/medien/publikation/long/4364.pdf

[72] "Fahrrad-monitor deutschland 2015: Ergebnisse einer repräsentativen online-befragung." [Online]. Available: http://www.bmvi.de/SharedDocs/DE/Anlage/VerkehrUndMobilitaet/Fahrrad/fahrrad-monitor-deutschland-2015.pdf?___blob=publicationFile

[73] K. Donaghy, G. Rudinger, and S. Poppelreuter, "Societal trends, mobility behaviour and sustainable transport in europe and north america," *Transport Reviews*, vol. 24, no. 6, pp. 679–690, 2004.

[74] L. Ding and N. Zhang, "A travel mode choice model using individual grouping based on cluster analysis," *Procedia Engineering*, vol. 137, pp. 786–795, 2016.

[75] C. Weis, M. Vrtic, P. Widmer, and K. W. Axhausen, "Influence of parking on location and mode choice, a stated choice survey."

[76] M. Braun Kohlová, "Everyday travel mode choice and its determinants: trip attributes versus lifestyle," Ph.D. dissertation, Charles University, Prague, 2016. [Online]. Available: www.feem-web.it/ess/files/braunkohlova.pdf

[77] B. Clark, K. Chatterjee, and S. Melia, "Changes to commute mode: The role of life events, spatial context and environmental attitude," *Transportation Research Part A: Policy and Practice*, vol. 89, pp. 89–105, 2016.

[78] H. M. Badland, N. Garrett, and G. M. Schofield, "How does car parking availability and public transport accessibility influence work-related travel behaviors?" *Sustainability*, vol. 2, no. 2, pp. 576–590, 2010.

[79] D. Banister and A. Bowling, "Quality of life for the elderly: The transport dimension," *Transport Policy*, vol. 11, no. 2, pp. 105–115, 2004.

[80] F. Sharmeen and H. Timmermans, "Walking down the habitual lane: Analyzing path dependence effects of mode choice for social trips," *Journal of Transport Geography*, vol. 39, pp. 222–227, 2014.

[81] L. Böcker, P. van Amen, and M. Helbich, "Elderly travel frequencies and transport mode choices in greater rotterdam, the netherlands," *Transportation*, 2016.

[82] M. J. Koetse and P. Rietveld, "The impact of climate change and weather on transport: An overview of empirical findings," *Transportation Research Part D: Transport and Environment*, vol. 14, no. 3, pp. 205–221, 2009.

[83] Y. Al Hassan and D. J. Barker, "The impact of unseasonable or extreme weather on traffic activity within lothian region, scotland," *Journal of Transport Geography*, vol. 7, no. 3, pp. 209–213, 1999.

[84] K. Keay and I. Simmonds, "The association of rainfall and other weather variables with road traffic volume in melbourne, australia," *Accident; analysis and prevention*, vol. 37, no. 1, pp. 109–124, 2005.

[85] M. Sabir, M. Koetse, and P. Rietveld, "The impact of weather conditions on mode choice: Empirical evidence for the netherlands," Ph.D. dissertation, Vrije University Amsterdam, Amsterdam, 2008. [Online]. Available: http://www.webmeets.com/files/papers/EAERE/2009/1021/Sabir.pdf

[86] M. Cools, E. Moons, and G. Wets, "Assessing the impact of weather on traffic intensity," *Weather, Climate, and Society*, vol. 2, no. 1, pp. 60–68, 2010.

[87] M. B. Kobus, E. Gutiérrez-i Puigarnau, P. Rietveld, and J. N. van Ommeren, "The on-street parking premium and car drivers' choice between street and garage parking," *Regional Science and Urban Economics*, vol. 43, no. 2, pp. 395–403, 2013.

[88] J. Yue, T. Cheng, and M. Tai, "Demand forecasting of parking lot based on discrete choice model in planned special events," in *2009 International Conference on Management and Service Science (MASS)*, 2009, pp. 1–4.

[89] A. E. Papacharalampous, S. Hovelynck, O. Cats, J. W. Lankhaar, W. Daamen, N. van Oort, and J. van Lint, "Multi-modal data fusion for big events [research news]," *IEEE Intelligent Transportation Systems Magazine*, vol. 7, no. 4, pp. 5–10, 2015.

[90] J. Collura, D. Fisher, and A. Holton, "Advanced parking management systems: Strategic management of intermodal junctions." [Online]. Available: http://ntl.bts.gov/lib/20000/20700/20741/PB98144678.pdf

[91] K. Spitaels and S. Maerivoet, "An empirical agent-based model of parking behaviour," Ph.D. dissertation, Transport & Mobility Lueven, 2008. [Online]. Available: https://www.researchgate.net/publication/228771303_An_empirical_agent-based_model_of_parking_behaviour

[92] D. A. Tsamboulas, "Parking fare thresholds: A policy tool," *Transport Policy*, vol. 8, no. 2, pp. 115–124, 2001.

[93] J. Golias, G. Yannis, and M. Harvatis, "Off-street parking choice sensitivity," *Transportation Planning and Technology*, vol. 25, no. 4, pp. 333–348, 2002.

[94] S. An, B. Han, and J. Wang, “Study of the mode of real-time and dynamic parking guidance and information systems based on fuzzy clustering analysis,” in *2004 International Conference on Machine Learning and Cybernetics*, . 2004, pp. 2790–2794.

[95] N. C. Balijepalli, S. P. Shepherd, and A. D. May, “Modelling the choice of car parks in urban areas and managing the demand for parking,” Ph.D. dissertation, University of Leeds, 2008. [Online]. Available: http://eprints.whiterose.ac.uk/3688

[96] Z. Pu, Z. Li, J. Ash, W. Zhu, and Y. Wang, “Evaluation of spatial heterogeneity in the sensitivity of on-street parking occupancy to price change,” *Transportation Research Part C: Emerging Technologies*, vol. 77, pp. 67–79, 2017.

[97] J. Kelly and J. Clinch, “Influence of varied parking tariffs on parking occupancy levels by trip purpose,” *Transport Policy*, vol. 13, no. 6, pp. 487–495, 2006.

[98] F. Pereira, A. Bazzan, and M. Ben-Akiva, “The role of context in transport prediction,” *IEEE Intelligent Systems*, vol. 29, no. 1, pp. 76–80, 2014.

[99] T. Rajabioun and P. Ioannou, “On-street and off-street parking availability prediction using multivariate spatiotemporal models,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 5, pp. 2913–2924, 2015.

[100] C. Pflügler, T. Köhn, M. Schreieck, M. Wiesche, and H. Krcmar, “Predicting the availability of parking spaces with publicly available data,” Ph.D. dissertation, Gesellschaft für Informatik, Bonn, 2016. [Online]. Available: https://www.excell-mobility.de/wp-content/uploads/2016/10/Predicting-the-Availability-of-Parking-Spaces.pdf

[101] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee, “Measurement and analysis of online social networks,” in *7th ACM SIGCOMM conference*, C. Dovrolis and M. Roughan, Eds., 2007, p. 29.

[102] Z. Xu, Y. Liu, N. Yen, L. Mei, X. Luo, X. Wei, and C. Hu, “Crowdsourcing based description of urban emergency events using social media big data,” *IEEE Transactions on Cloud Computing*, p. 1, 2016.

[103] A. Candelieri and F. Archetti, “Detecting events and sentiment on twitter for improving urban mobility,” Ph.D. dissertation, University of Milano-Bicocca, 2015. [Online]. Available: http://ceur-ws.org/Vol-1351/paper9.pdf

[104] X. Sun, Y. Wu, L. Liu, and J. Panneerselvam, “Efficient event detection in social media data streams,” in *2015 IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communica-*

tions; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing (CIT/IUCC/DASC/PICOM), 2015, pp. 1711–1717.

[105] Z. Dashdorj, B. Tsogtbaatar, A. Tumurchudur, and E. Altangerel, "High level event identification in social media," in *2016 12th International Conference on Semantics, Knowledge and Grids (SKG)*, 2016, pp. 121–125.

[106] N. Alsaedi, P. Burnap, and O. Rana, "Sensing real-world events using social media data and a classification-clustering framework," in *2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, 2016, pp. 216–223.

[107] L. Barkhuus and J. Tashiro, "Student socialization in the age of facebook," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, E. Mynatt, D. Schoner, G. Fitzpatrick, S. Hudson, K. Edwards, and T. Rodden, Eds., 2010, p. 133.

[108] L. Sullivan, "Social media interactions may influence offline behavior," 2012. [Online]. Available: http://www.mediapost.com/publications/article/184044/social-media-interactions-may-influence-offline-be.html

[109] S. Vissers, M. Hooghe, D. Stolle, and V.-A. Maheo, "The impact of mobilization media on off-line and online participation: Are mobilization effects medium-specific?" *Social Science Computer Review*, vol. 30, no. 2, pp. 152–169, 2012.

[110] R. Du, Z. Yu, T. Mei, Z. Wang, Z. Wang, and B. Guo, "Predicting activity attendance in event-based social networks," in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing - UbiComp '14 Adjunct*, A. J. Brush, A. Friday, J. Kientz, J. Scott, and J. Song, Eds. New York, New York, USA: ACM Press, 2014, pp. 425–434.

[111] D. Nguyen and T. Le, "Recommendation system for facebook public events based on probabilistic classification and re-ranking," in *2016 Eighth International Conference on Knowledge and Systems Engineering (KSE)*. IEEE, 2016, pp. 133–138.

[112] M. Bogaert, M. Ballings, and D. van den Poel, "The added value of facebook friends data in event attendance prediction," *Decision Support Systems*, vol. 82, pp. 26–34, 2016.

[113] J. Han, J. Niu, A. Chin, W. Wang, C. Tong, and X. Wang, "How online social network affects offline events: A case study on douban," in *2012 9th International Conference on Ubiquitous Intelligence and Computing and 9th International Conference on Autonomic and Trusted Computing*. IEEE, 2012, pp. 752–757.

[114] M. Kawano, T. Yonezawa, J. Nakazawa, S. Kawasaki, K. Ohta, H. Inamura, and H. Tokuda, "Classifying urban events' popularity by analyzing friends information in location-based social network," in *The First International Conference on IoT in Urban Space*, F. Kawsar, U. Blanke, A. Mashhadi, and B. Altakrouri, Eds., 2014.

[115] E. Mynatt and J. Tullio, "Inferring calendar event attendance," in *Proceedings of the 6th international conference on Intelligent user interfaces*, C. Sidner and J. Moore, Eds., 2001, pp. 121–128.

[116] C. M. Paris, W. Lee, and P. Seery, "The role of social media in promoting special events: Acceptance of facebook 'events'," in *Information and Communication Technologies in Tourism 2010*, U. Gretzel, R. Law, and M. Fuchs, Eds. Vienna: Springer Vienna, 2010, pp. 531–541.

[117] W. Lee, L. Xiong, and C. Hu, "The effect of facebook users' arousal and valence on intention to go to the festival: Applying an extension of the technology acceptance model," *International Journal of Hospitality Management*, vol. 31, no. 3, pp. 819–827, 2012.

[118] J. Michalco and P. Navrat, "Arrangement of face-to-face meetings using social media," Ph.D. dissertation, 2012. [Online]. Available: https://www.researchgate.net/publication/241432335_Arrangement_of_Face-to-Face_Meetings_Using_Social_Media

[119] A.-J. Huang, H.-C. Wang, and C. W. Yuan, "De-virtualizing social events: Understanding the gap between online and offline participation for event invitations," in *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, S. Fussell, W. Lutters, M. R. Morris, and M. Reddy, Eds., 2014, pp. 436–448.

[120] "Facebook's global economic impact." [Online]. Available: https://www2.deloitte.com/content/dam/Deloitte/uk/Documents/technology-media-telecommunications/deloitte-uk-global-economic-impact-of-facebook.pdf

[121] McMahon and Paul, "Lessons learned from hosting over 10,000 events," 2016. [Online]. Available: https://www.doorkeeperhq.com/event-planning/increasing-participants-decreasing-no-shows

[122] X. Lu, E. Wetter, N. Bharti, A. J. Tatem, and L. Bengtsson, "Approaching the limit of predictability in human mobility," *Scientific reports*, vol. 3, p. 2923, 2013.

[123] F. C. Pereira, F. Rodrigues, and M. Ben-Akiva, "Using data from the web to predict public transport arrivals under special events scenarios," *Journal of Intelligent Transportation Systems*, vol. 19, no. 3, pp. 273–288, 2014.

[124] A. David and H. Keller, "Event-driven modelling of on-street parking probability," 2001 ITS World Congress in Sydney. [Online]. Available: http://irandanesh.febpco.com/FileEssay/barnamerizi-1386-12-8-bgh(176).PDF

[125] "Mobinet abschlussbericht 2003," München. [Online]. Available: http://www.mvv-muenchen.de/fileadmin/media/Dateien/7_Der_MVV/dokumente/Mobinet_Abschlussbericht_2003_gesamt.pdf

[126] Y. Li, M. Steiner, L. Wang, Z.-L. Zhang, and J. Bao, "Exploring venue popularity in foursquare," in *2013 Proceedings IEEE INFOCOM*. IEEE, 2013, pp. 3357–3362.

[127] A. S. Furtado, R. Fileto, and C. Renso, "M-attract: Assessing the attractiveness of places by using moving objects trajectories data," in *Proceedings of the 13th GEOINFO*, 2012, pp. 84–95. [Online]. Available: http://www.geoinfo.info/proceedings/proceedings_geoinfo2012.pdf#page=91

[128] J. Sun and R. Wang, "The study of parking demand forecast basing on the analysis of the land location," in *2011 International Conference on Remote Sensing, Environment and Transportation Engineering (RSETE)*, 2011, pp. 2752–2755.

[129] T. Giuffrè, S. M. Siniscalchi, and G. Tesoriere, "A novel architecture of parking management for smart cities," *Procedia - Social and Behavioral Sciences*, vol. 53, pp. 16–28, 2012.

[130] D. Landry and M. Morin, "Estimating parking spot occupancy," Ph.D. dissertation, Brigham Young University, 2013. [Online]. Available: http://www.et.byu.edu/~bmazzeo/ECEn_670_F13/mini_conference_files/Morin_Landry_paper.pdf

[131] Y.-j. Ji, D.-n. Tang, W.-h. Guo, P. T. Blythe, and W. Wang, "Forecasting available parking space with largest lyapunov exponents method," *Journal of Central South University*, vol. 21, no. 4, pp. 1624–1632, 2014.

[132] P. Blythe, Y. Ji, W. Guo, W. Wang, and D. Tang, "Short-term forecasting of available parking space using wavelet neural network model," *IET Intelligent Transport Systems*, vol. 9, no. 2, pp. 202–209, 2015.

[133] F. Yu, J. Guo, X. Zhu, and G. Shi, "Real time prediction of unoccupied parking space using time series model," in *2015 International Conference on Transportation Information and Safety (ICTIS)*, 2015, pp. 370–374.

[134] M. Caliskan, A. Barthels, B. Scheuermann, and M. Mauve, "Predicting parking lot occupancy in vehicular ad hoc networks," in *2007 IEEE 65th Vehicular Technology Conference*, 2007, pp. 277–281.

[135] A. Ziat, B. Leroy, N. Baskiotis, and L. Denoyer, "Joint prediction of road-traffic and parking occupancy over a city with representation learning," in *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*, 2016, pp. 725–730.

[136] L. Chang, "Development of prediction model for real-time parking availability for on-street paid parking," Master Thesis, University of Pittsburgh, 2016. [Online]. Available: http://d-scholarship.pitt.edu/id/eprint/26353

[137] K. Martens and I. Benenson, "Evaluating urban parking policies with agent-based model of driver parking behavior," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2046, pp. 37–44, 2008.

[138] F. Caicedo, C. Blazquez, and P. Miranda, "Prediction of parking space availability in real time," *Expert Systems with Applications*, vol. 39, no. 8, pp. 7281–7290, 2012.

[139] S. Yong, L. Chunping, W. Yihuai, Z. Shukui, and L. Feixiong, "A forecasting model for parking guidance system," in *2009 WRI World Congress on Computer Science and Information Engineering*, 2009, pp. 607–611.

[140] Z. Yang, H. Liu, and X. Wang, "The research on the key technologies for improving efficiency of parking guidance system," in *2003 IEEE International Conference on Intelligent Transportation Systems*, 2003, pp. 1177–1182.

[141] X. Chen, "Parking occupancy prediction and pattern analysis," Ph.D. dissertation, 2014. [Online]. Available: http://cs229.stanford.edu/proj2014/Xiao%20Chen,Parking%20Occupancy%20Prediction%20and%20Pattern%20Analysis.pdf

[142] F. Richter, S. Di Martino, and D. C. Mattfeld, "Temporal and spatial clustering for a parking prediction service," in *2014 IEEE 26th International Conference on Tools with Artificial Intelligence (ICTAI)*, 2014, pp. 278–282.

[143] E. I. Vlahogianni, K. Kepaptsoglou, V. Tsetsos, and M. G. Karlaftis, "A real-time parking prediction system for smart cities," *Journal of Intelligent Transportation Systems*, vol. 20, no. 2, pp. 192–204, 2015.

[144] "Facebook q4 2016 results," 2017. [Online]. Available: https://s21.q4cdn.com/399680738/files/doc_presentations/FB-Q4'16-Earnings-Slides.pdf

[145] R. Lawler, "Eventbrite now recommends events to 20 million users, boosting ticket sales along the way," 2012. [Online]. Available: https://techcrunch.com/2012/08/03/eventbrite-recommendations/

[146] "Traffic statistics," 2017. [Online]. Available: http://www.alexa.com/siteinfo/

[147] "Percentage of global population using facebook as of june 2016, by region," 2016. [Online]. Available: https://www.statista.com/statistics/241552/share-of-global-population-using-facebook-by-region/

[148] "Net digital advertising revenue share of major ad-selling online companies worldwide from 2016 to 2019," 2016. [Online]. Available: https://www.statista.com/statistics/290629/digital-ad-revenue-share-of-major-ad-selling-companies-worldwide/

[149] R. Wilken, "Places nearby: Facebook as a location-based social media platform," *New Media & Society*, vol. 16, no. 7, pp. 1087–1103, 2014.

[150] "Factual and facebook expand location data partnership," 2016. [Online]. Available: https://www.factual.com/blog/factual-and-facebook-expand-location-data-partnership

[151] "Anzahl der registrierten nutzer von foursquare weltweit in ausgewählten monaten von dezember 2010 bis januar 2015," 2015. [Online]. Available: https://de.statista.com/statistik/daten/studie/302990/umfrage/anzahl-der-nutzer-von-foursquare-weltweit/

[152] P. Neis, "Statistics of the free wiki world map (openstreetmap.org)," 2017. [Online]. Available: http://osmstats.neis-one.org/

[153] W. Castelein, L. Grus, J. Crompvoets, and A. Bregt, "A characterization of volunteered geographic information," in *13th AGILE International Conference on Geographic Information Science*, 2010. [Online]. Available: https://agile-online.org/conference_paper/cds/agile_2010/shortpapers_pdf/106_doc.pdf

[154] M. Casserly, "Eventbrite brags $2 billion in ticket sales: $500 million in just six months," 2013. [Online]. Available: https://www.forbes.com/sites/meghancasserly/2013/09/25/eventbrite-brags-2-billion-in-ticket-sales-500-million-in-just-six-months/#b4aa4bd6e032

[155] "Welcome to the fan-centric platform." 2017. [Online]. Available: https://developer.ticketmaster.com/

[156] "Live nation reports $7.2 billion in revenue for 2015," 2016. [Online]. Available: http://www.billboard.com/articles/business/6890347/live-nation-2015-earnings-7-6-billion-revenue

[157] "Databank - world development indicators: Urban population (% of total)," 2017. [Online]. Available: http://databank.worldbank.org/data/reports.aspx?source=2&series=SP.URB.TOTL.IN.ZS&country=DEU

[158] "List of urban areas by country," 2012. [Online]. Available: http://www.oecd.org/cfe/regional-policy/all.pdf

[159] "Planet dump," 2017. [Online]. Available: https://planet.osm.org

[160] "Discovery api: v. 2.0," 2017. [Online]. Available: http://developer.ticketmaster.com/products-and-docs/apis/discovery-api/v2/

[161] Y. Zhang, H. Wu, A. Panangadan, and V. K. Prasanna, "Integration of heterogeneous web services for event-based social networks," in *2015 IEEE International Conference on Information Reuse and Integration.* IEEE, 2015, pp. 57–63.

[162] G. Recchia and M. Louwerse, "A comparison of string similarity measures for toponym matching," in *Proceedings of The First ACM SIGSPATIAL International Workshop on Computational Models of Place - COMP '13.* New York, New York, USA: ACM Press, 2013, pp. 54–61.

[163] Python Software Foundation, "difflib: Helpers for computing deltas," 2017. [Online]. Available: https://docs.python.org/3.6/library/difflib.html

[164] "Wordnet: A lexical databse for english," 2017. [Online]. Available: https://wordnet.princeton.edu/

[165] C. M. Bishop, *Pattern recognition and machine learning*, corrected at 8. printing 2009 ed., ser. Information science and statistics. New York, NY: Springer, 2009.

[166] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[167] R. L. Figueroa, Q. Zeng-Treitler, S. Kandula, and L. H. Ngo, "Predicting sample size required for classification performance," *BMC medical informatics and decision making*, vol. 12, p. 8, 2012.

[168] N. Boonyanunta and P. Zeephongsekul, "Predicting the relationship between the size of training sample and the predictive power of classifiers," in *Knowledge-Based Intelligent Information and Engineering Systems*, ser. Lecture Notes in Computer Science, D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, M. G. Negoita, R. J. Howlett, and L. C. Jain, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, vol. 3215, pp. 529–535.

[169] "Interaktiver regionalatlas," 2017. [Online]. Available: http://www.statistik-portal.de/Statistik-Portal/Regionalatlas/

Appendix

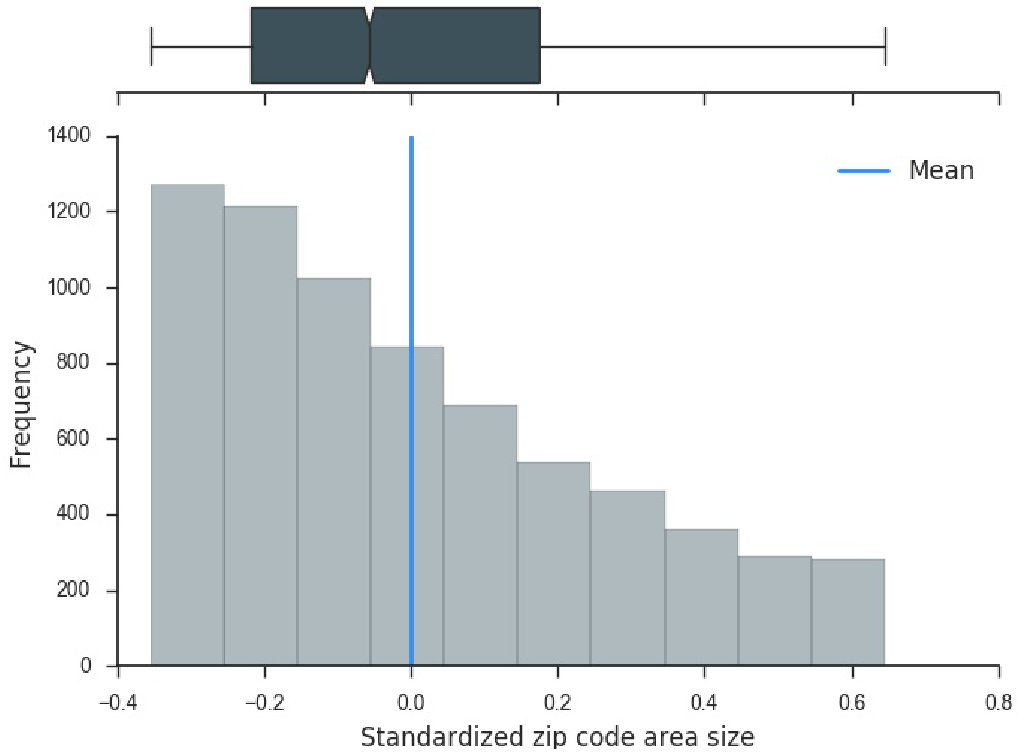


Figure F1. Distribution of zip code areas in Germany by size

Excluded Facebook place categories	
City	Continent
Region	Volcano
Postal Code	Bay
Country	Field
State/Province/Region	Province
Ocean	

Table T1. Manually excluded place categories on Facebook

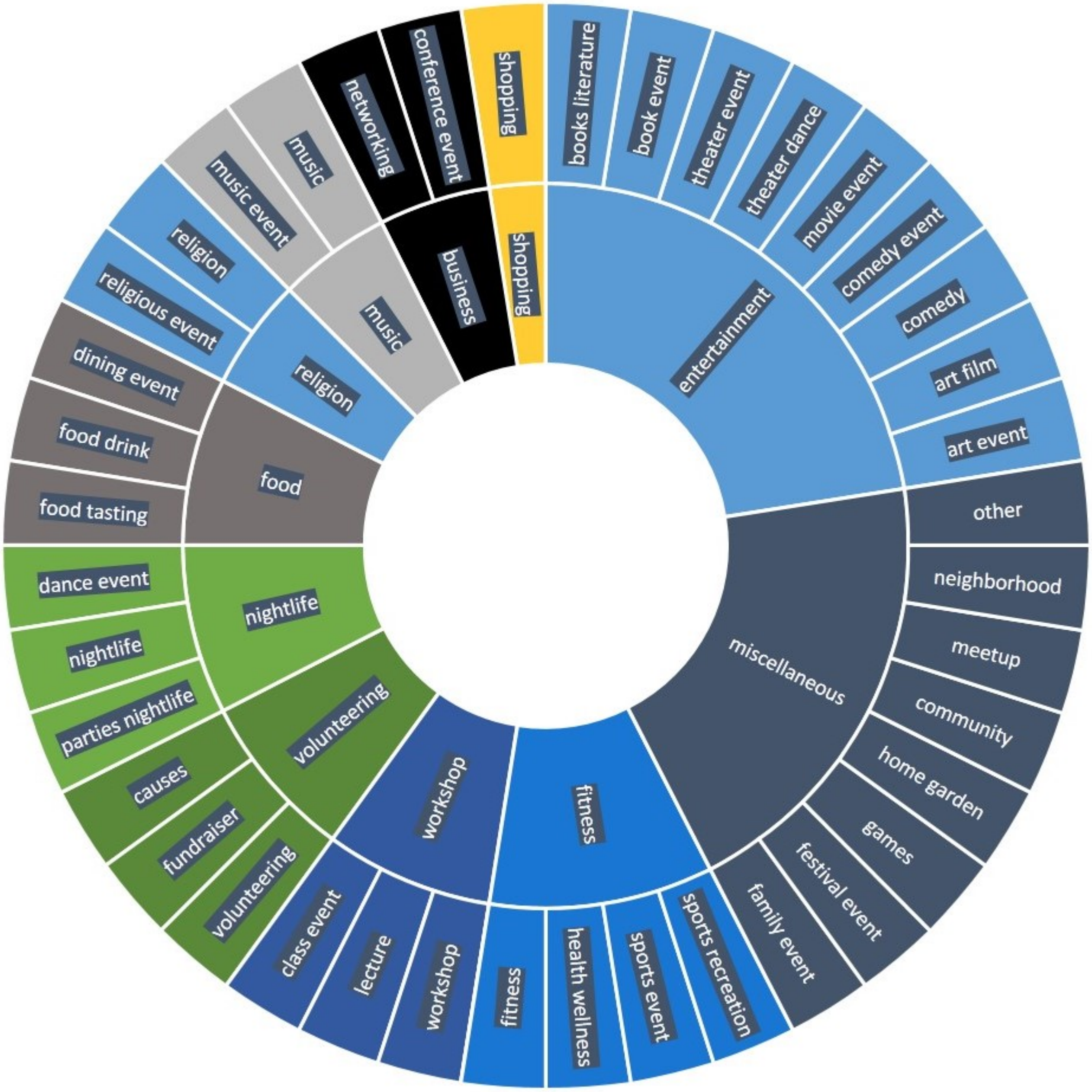


Figure F2. Assignment of Facebook event categories to superior segments

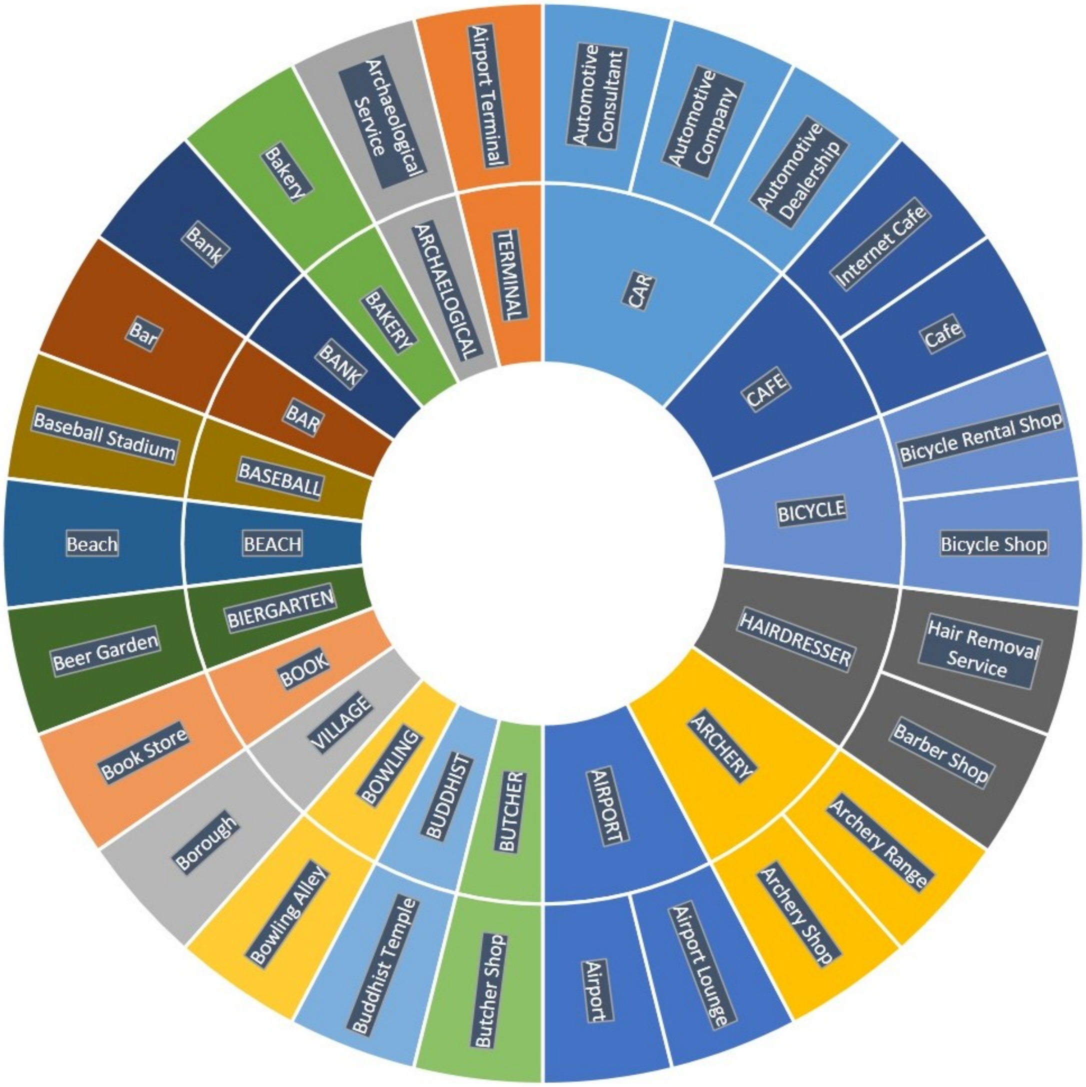


Figure F3. Sample assignment of Facebook and OSM categories regarding POIs

Ticketmaster	Facebook	Eventbrite
Arts & Theatre	Art Event	Performing & Visual Arts
Arts & Theatre	Art Film	Film, Media & Entertainment
	Book Event	Hobbies & Special Interest
	Books Literature	Hobbies & Special Interest
	Causes	Charity & Causes
	Class Event	Family & Education
	Comedy	Film, Media & Entertainment
	Comedy Event	Film, Media & Entertainment
	Community	Community & Culture
	Conference Event	Business & Professional
	Dance Event	
	Dining Event	Food & Drink
	Family Event	Family & Education
	Festival Event	Seasonal & Holiday
	Fitness	Sports & Fitness
	Food Drink	Food & Drink
	Food Tasting	Food & Drink
	Fundraiser	Charity & Causes
	Games	Hobbies & Special Interest
	Health Wellness	Health & Wellness
	Home Garden	Home & Lifestyle
	Meetup	Community & Culture
Film	Movie Event	Film, Media & Entertainment
Music	Music	Music
Music	Music Event	Music
	Neighborhood	Community & Culture
	Networking	Business & Professional
	Nightlife	
	Parties Nightlife	
	Religion	Religion & Spirituality
	Religious Event	Religion & Spirituality
	Shopping	
Sports	Sports Event	Sports & Fitness
Sports	Sports Recreation	Sports & Fitness
Arts & Theatre	Theater Dance	Performing & Visual Arts
Arts & Theatre	Theater Event	Performing & Visual Arts
	Volunteering	Charity & Causes
	Workshop	Family & Education

Table T2. Categorical matching results for event data sources

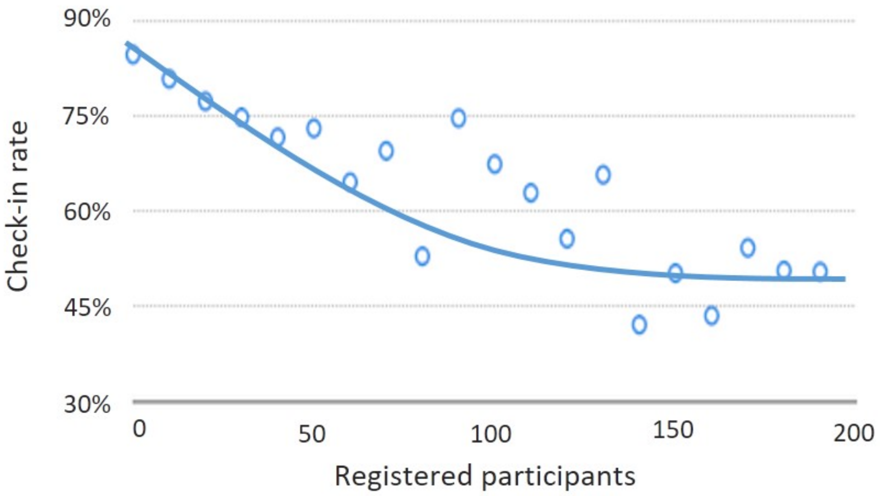


Figure F4. Trend line for check-in rate by event size on Doorkeeper platform; based on [121]

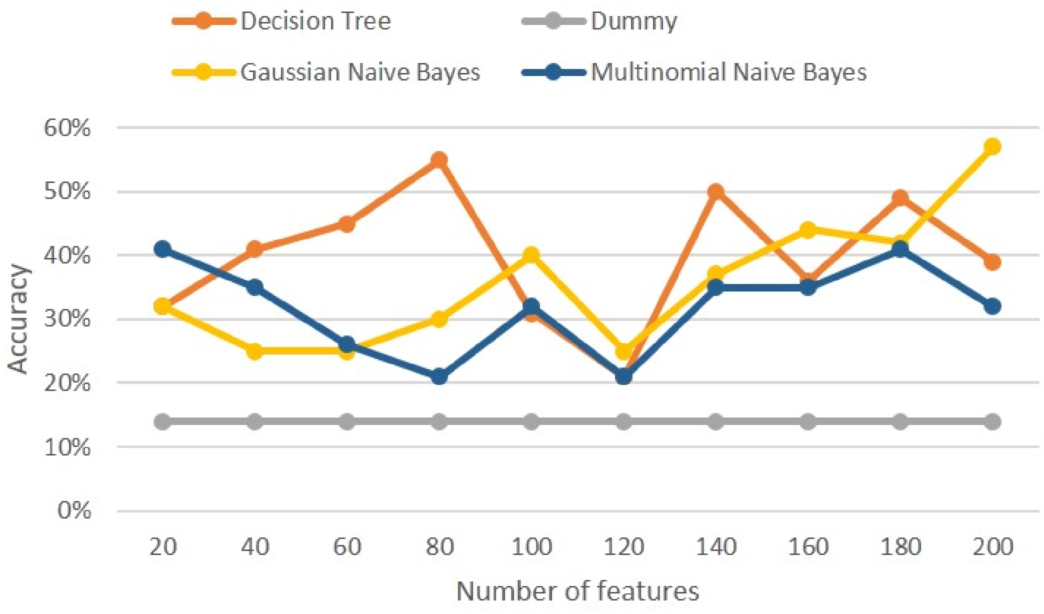


Figure F5. Classification accuracy for alcohol attribute

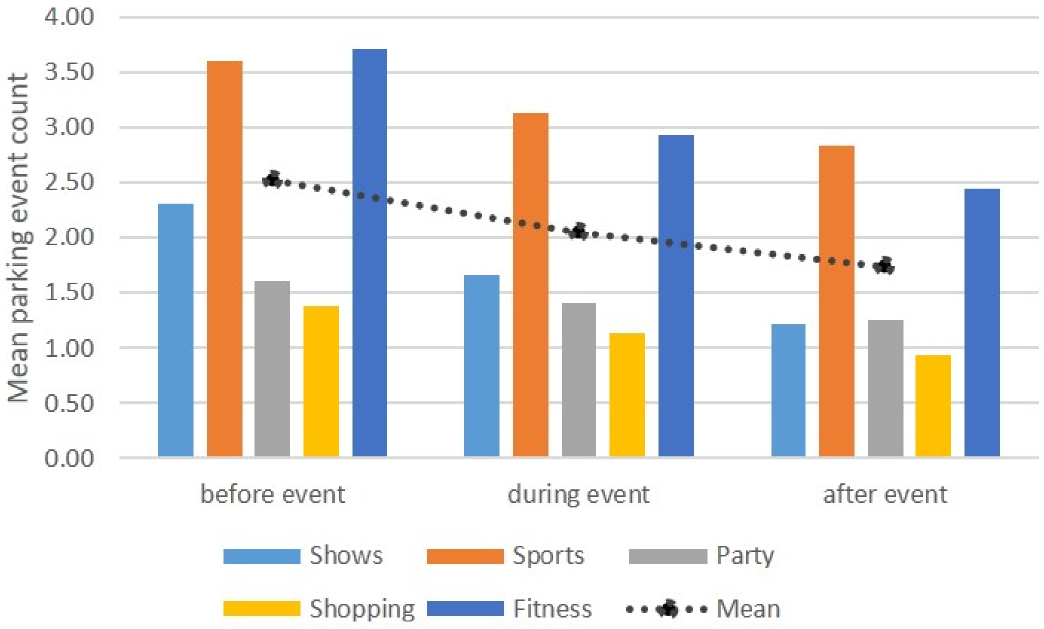


Figure F6. Probabilistic parking event distribution per topic (2.0 km focus polygons)

Library	Version
distance	0.1.3
scikit-learn	0.18.1
pandas	0.19.2
folium	0.3.0
seaborn	0.7.1
joblib	0.9.4
numpy	1.11.3
matplotlib	1.5.3
requests	2.13.0
treetaggerwrapper	2.2.4
spyder	3.1.3
nltk	3.2.1
anaconda	4.3.16
ipython	5.1.0

Table T3. Python packages used

BIBLIOGRAPHY

"Facebook's global economic impact." [Online]. Available: https://www2.deloitte.com/content/dam/Deloitte/uk/Documents/technology-media-telecommunications/deloitte-uk-global-economic-impact-of-facebook.pdf

"Parken 2020: Quantum fokus." [Online]. Available: https://www.quantum.ag/fileadmin/Dateien/Publikationen_Archiv/QuantumFokus_2-2012.pdf

"List of urban areas by country," 2012. [Online]. Available: http://www.oecd.org/cfe/regional-policy/all.pdf

"Anzahl der registrierten nutzer von foursquare weltweit in ausgewählten monaten von dezember 2010 bis januar 2015," 2015. [Online]. Available: https://de.statista.com/statistik/daten/studie/302990/umfrage/anzahl-der-nutzer-von-foursquare-weltweit/

"Smart parking: A system that could help cities rethink parking: Background information," 2015. [Online]. Available: http://www.siemens.com/press/pool/de/events/2015/corporate/2015-09-iaa/background-smart-parking-e.pdf

"Factual and facebook expand location data partnership," 2016. [Online]. Available: https://www.factual.com/blog/factual-and-facebook-expand-location-data-partnership

"Live nation reports $7.2 billion in revenue for 2015," 2016. [Online]. Available: http://www.billboard.com/articles/business/6890347/live-nation-2015-earnings-7-6-billion-revenue

"Net digital advertising revenue share of major ad-selling online companies worldwide from 2016 to 2019," 2016. [Online]. Available: https://www.statista.com/statistics/290629/digital-ad-revenue-share-of-major-ad-selling-companies-worldwide/

"Percentage of global population using facebook as of june 2016, by region," 2016. [Online]. Available: https://www.statista.com/statistics/241552/share-of-global-population-using-facebook-by-region/

"Databank - world development indicators: Urban population (% of total)," 2017. [Online]. Available: http://databank.worldbank.org/data/reports.aspx?source=2&series=SP.URB.TOTL.IN.ZS&country=DEU

"Discovery api: v. 2.0," 2017. [Online]. Available: http://developer.ticketmaster.com/products-and-docs/apis/discovery-api/v2/

"Facebook q4 2016 results," 2017. [Online]. Available: https://s21.q4cdn.com/399680738/files/doc_presentations/FB-Q4'16-Earnings-Slides.pdf

"Interaktiver regionalatlas," 2017. [Online]. Available: http://www.statistik-portal.de/Statistik-Portal/Regionalatlas/

"Planet dump," 2017. [Online]. Available: https://planet.osm.org

"Traffic statistics," 2017. [Online]. Available: http://www.alexa.com/siteinfo/

"Welcome to the fan-centric platform." 2017. [Online]. Available: https://developer.ticketmaster.com/

"Wordnet: A lexical databse for english," 2017. [Online]. Available: https://wordnet.princeton.edu/

Belloche, S., "On-street parking search time modelling and validation with survey-based data," *Transportation Research Procedia*, vol. 6, pp. 313–324, 2015.

Bishop, C. M., *Pattern recognition and machine learning*, corrected at 8. printing 2009 ed., ser. Information science and statistics. New York, NY: Springer, 2009.

Boonyanunta, N. and Zeephongsekul, P., "Predicting the relationship between the size of training sample and the predictive power of classifiers," in *Knowledge-Based Intelligent Information and Engineering Systems*, ser. Lecture Notes in Computer Science, Hutchison, D., Kanade, T., Kittler, J., Kleinberg, J. M., Mattern, F., Mitchell, J. C., Naor, M., Nierstrasz, O., Pandu Rangan, C., Steffen, B., Sudan, M., Terzopoulos, D., Tygar, D., Vardi, M. Y., Weikum, G., Negoita, M. G., Howlett, R. J., and Jain, L. C., Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, vol. 3215, pp. 529–535.

Casserly, M., "Eventbrite brags $2 billion in ticket sales: $500 million in just six months," 2013. [Online]. Available: https://www.forbes.com/sites/meghancasserly/2013/09/25/eventbrite-brags-2-billion-in-ticket-sales-500-million-in-just-six-months/#b4aa4bd6e032

Castelein, W., Grus, L., Crompvoets, J., and Bregt, A., "A characterization of volunteered geographic information," in *13th AGILE International Conference on Geographic Information Science*, 2010. [Online]. Available: https://agile-online.org/conference_paper/cds/agile_2010/shortpapers_pdf/106_doc.pdf

Chen, X. and Liu, N., "Smart parking by mobile crowdsensing," *International Journal of Smart Home*, vol. 10, no. 2, pp. 219–234, 2016.

Figueroa, R. L., Zeng-Treitler, Q., Kandula, S., and Ngo, L. H., “Predicting sample size required for classification performance,” *BMC medical informatics and decision making*, vol. 12, p. 8, 2012.

Huang, A.-J., Wang, H.-C., and Yuan, C. W., “De-virtualizing social events: Understanding the gap between online and offline participation for event invitations,” in *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, Fussell, S., Lutters, W., Morris, M. R., and Reddy, M., Eds., 2014, pp. 436–448.

Lawler, R., “Eventbrite now recommends events to 20 million users, boosting ticket sales along the way,” 2012. [Online]. Available: https://techcrunch.com/2012/08/03/eventbrite-recommendations/

McMahon and Paul, “Lessons learned from hosting over 10,000 events,” 2016. [Online]. Available: https://www.doorkeeperhq.com/event-planning/increasing-participants-decreasing-no-shows

Ministry of Public Security China and Xinhua News Agency, “Vehicle population in china from 2007 to 2015,” 2016. [Online]. Available: https://www.statista.com/statistics/285306/number-of-car-owners-in-china/

Morillo, C. and Campos, J. M., “On-street illegal parking costs in urban areas,” *Procedia - Social and Behavioral Sciences*, vol. 160, pp. 342–351, 2014.

Neis, P., “Statistics of the free wiki world map (openstreetmap.org),” 2017. [Online]. Available: http://osmstats.neis-one.org/

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E., “Scikit-learn: Machine learning in python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

Pfoser, D., “Floating car data,” in *Encyclopedia of GIS*, Shekhar, S. and Xiong, H., Eds. Boston, MA: Springer US, 2008, p. 321.

Python Software Foundation, “difflib: Helpers for computing deltas,” 2017. [Online]. Available: https://docs.python.org/3.6/library/difflib.html

Recchia, G. and Louwerse, M., “A comparison of string similarity measures for toponym matching,” in *Proceedings of The First ACM SIGSPATIAL International Workshop on Computational Models of Place - COMP ’13*. New York, New York, USA: ACM Press, 2013, pp. 54–61.

Shoup, D. C., “Cruising for parking,” *Transport Policy*, vol. 13, no. 6, pp. 479–486, 2006.

Wilken, R., “Places nearby: Facebook as a location-based social media platform,” *New Media & Society*, vol. 16, no. 7, pp. 1087–1103, 2014.

Zhang, H. and Malczewski, J., “Quality evaluation of volunteered geographic information,” in *Volunteered Geographic Information and the Future of Geospatial Data*, ser. Advances in Geospatial Technologies, Dey, N., Calazans Campelo, C. E., Bertolotto, M., and Corcoran, P., Eds. IGI Global, 2017, pp. 19–46.

Zhang, Y., Wu, H., Panangadan, A., and Prasanna, V. K., “Integration of heterogeneous web services for event-based social networks,” in *2015 IEEE International Conference on Information Reuse and Integration*. IEEE, 2015, pp. 57–63.