Technische Universität München
TUM School of Engineering and Design

Developing a Methodology for Quality Assessment of Mobility-Related Information Systems for Connected Vehicles:

The Case of On-Street Parking

Syrus Borja Gomari

Vollständiger Abdruck der von der TUM School of Engineering and Design der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Ingenieurwissenschaften

genehmigten Dissertation.

Vorsitz:
Prof. Dr.-Ing. Rolf Moeckel

Prüfer der Dissertation:
1. Prof. Dr. Constantinos Antoniou
2. Prof. Dr.-Ing. Klaus Bogenberger

Die Dissertation wurde am 21.04.2022 bei der Technischen Universität München eingereicht und durch TUM School of Engineering and Design am 25.10.2022 angenommen.

Acknowledgements

This dissertation would not have been completed without the guidance and support of my supervisor, Professor Dr. Constantinos Antoniou. Throughout the entire journey, he provided invaluable scientific advice that shaped the work done in this research. I would also like to extend my gratitude to him as he constantly reminded me and helped me find ways to continuously be engaged in the academic environment through the university whilst pursuing this research together with an industry partner.

I would also like to express my gratitude to my mentor Dr. Christoph Knoth. My journey as a Doctoral Researcher at BMW Group started with him as my supervisor within the company's Connected Parking team. I am thankful for all the feedback and stimulating discussions with him, which were instrumental in identifying the scientific research questions that were answered in this dissertation. I will forever cherish the collaborative work we had done within the BMW Group, which daily reminded me of the importance of also conducting applied research.

I want to acknowledge the BMW Group as one of the main enablers of this research work. The opportunity to conduct research within a prestigious company in cooperation with a world-class university, that is, the Technical University of Munich, is something that I will eternally be grateful for. I am thankful to all those with whom I have had the pleasure to work within the BMW Group. Particularly, I want to express my gratitude to my team leader, Benjamin Krebs, who played a significant role in the latter half of my time at the company. His constant push for me to deliver the best work that I can within the team, despite the challenges that came in the last year, made me a better researcher.

I am also grateful to all my former colleagues and friends at the Chair of Transportation Systems Engineering. The formal and informal discussions helped me discover new methods to approach the research questions I faced in this journey.

I am deeply grateful to all the friends I made throughout the duration of my PhD. The support system I was able to build in Munich with a group of amazing friends always pushed me to my limits as a person and as a researcher.

Last but not the least, I would like to give my deepest gratitude to my siblings Laleh, Parviz, and Peiman, to my brother-in-law Stefan Luksch, to my father Safar Ali, and to my girlfriend Charmaine Guzman. Without their endless support and understanding through the ups and downs of the entire journey, this work would not have been possible. For that, I will forever be indebted to them.

All this work is dedicated in the memory of my mother, Divina Borja Gomari.

Abstract

Mobility-related information systems, particularly, the use case of on-street parking information (OSPI) systems have increased in popularity as it lessens on-street parking search time and reduces congestion. OSPI service is one of the core elements of smarter navigation. A major challenge with the existing information systems is the proper assessment of such systems. There is a lack of scalable solutions that reduce reliability on manually collected ground truth data. A robust assessment methodology is needed to ensure high quality and reliable information is delivered to users. Three studies are consolidated within this dissertation which deal with the development of a quality assessment methodology.

Vehicle parked-in and parked-out events were utilised as the main data source to infer parking behaviour, to develop a prioritization-based quality assessment methodology, and to create a data-driven parking prediction model. These three components led to the development of a novel methodology which lessens the dependence on manually collected ground truth by carefully selecting areas that need to be observed, considered to be the important areas based on the frequency of visits as inferred from the parking events data. In line with that, a prioritization-based approach giving more weight to important areas adjusts the quality evaluation scores that are reflective of users' demand. Furthermore, with a comprehensive cluster analysis of parking behaviour, data-driven models were developed based on enhanced features that are able to keep the models up-to-date.

This dissertation lays the framework towards more comprehensive quality assessment methodologies necessary to ensure product quality of mobility-related information systems. Recommendations on further research directions on all the aspects of quality assessment are provided at the end of this document.

Zusammenfassung

Mobilitätsbezogene Informationssysteme, insbesondere On-Street-Parking-Informationssysteme (OSPI), erfreuen sich zunehmender Beliebtheit, da sie die Zeit für die Parkplatzsuche auf der Straße verkürzen und Staus reduzieren. Der OSPI-Dienst ist eines der Kernelemente einer intelligenteren Navigation. Eine große Herausforderung bei den gegenwärtig existierenden Informationssystemen ist die angemessene Bewertung dieser Systeme. Insbesondere mangelt es an skalierbaren Lösungen, die die Abhängigkeit von manuell erfassten Daten verringern. Um sicherzustellen, dass den Nutzern qualitativ hochwertige und zuverlässige Informationen zur Verfügung gestellt werden, ist eine robuste Bewertungsmethodik erforderlich. In dieser Dissertation werden drei Studien zusammengefasst, die sich mit der Entwicklung einer Qualitätsbewertungsmethodik befassen.

Ein- und Ausparkvorgänge wurden als Hauptdatenquelle genutzt, um Rückschlüsse auf das Parkverhalten zu ziehen, eine auf Prioritäten basierende Qualitätsbewertungsmethodik zu entwickeln und ein datengestütztes Parkvorhersagemodell zu erstellen. Diese drei Komponenten dieser Dissertation führten zur Entwicklung einer neuartigen Methodik, die die Abhängigkeit von manuell gesammeltem Ground Truth durch eine sorgfältige Auswahl der zu beobachtenden Bereiche, die aufgrund der aus den Parkereignisdaten abgeleiteten Häufigkeit der Besuche als wichtig erachtet werden, verringern kann. Ein auf Priorisierung basierender Ansatz, der wichtige Bereiche stärker gewichtet, passt die Qualitätsbewertungspunkte an, die die Nachfrage der Nutzer widerspiegeln. Darüber hinaus wurden mit einer umfassenden Clusteranalyse des Parkverhaltens datengesteuerte Modelle auf der Grundlage erweiterter Funktionen entwickelt, die die Modelle auf dem neuesten Stand halten können.

Diese Dissertation legt den Rahmen für umfassendere Qualitätsbewertungsmethoden fest, die notwendig sind, um die Produktqualität von mobilitätsbezogenen Informationssystemen zu gewährleisten. Empfehlungen für weitere Forschungsrichtungen zu allen Aspekten der Qualitätsbewertung werden am Ende dieses Dokuments gegeben.

Contents

List of Figures

List of Tables

1 Introduction

Emerging technologies have increasingly become a necessity in the mobility and transportation space. Within the last 15 years, society has experienced the emergence of on-demand ride-hailing companies, the development of easily accessible smartphone-based mobility-related information systems, and the surprising growth of the sharing economy (e.g., car sharing, ride sharing, bike sharing, and scooter sharing systems). Due to their rapid development, the quality of these systems remains questionable as new forms and demands emerge. Quality assurance becomes even more crucial as the inevitable and impending technological transition to more connected and automated vehicles gathers pace. Information systems will play a vital role for "talking" vehicles and users of the system. Poor quality or misguided information may be as bad as or worse than not providing any information as it could lead to unnecessary delays and conflicts in a system. This is more apparent in on-street parking information (OSPI) systems that have been around for over 5 years now. Inaccurate guides lead to circling around a neighbourhood and unnecessarily cause delays, which result in potential user losses of the system, exacerbation of traffic congestion, and deferring potential customers to businesses in the city.

The existing OSPI approach does not scale. Concurrently, the quality demanded by customers is increasing. Thus, the main problem this research aims to address is what methods need to be developed and used to efficiently and truthfully assess the quality of OSPI systems and support service rollouts with optimum quality.

This dissertation summarizes the author's research studies [1]–[4] and developments towards a better quality assessment methodology for mobility-related information systems, and particularly the use case of OSPI services.

1.1 Smarter Navigation: On-Street Parking Information Systems Background

In-car navigation systems have already existed for a while. However, after its initial roll out, providers and automotive Original Equipment Manufacturers (OEMs) realized end-of-trip pain points of customers when they are about to search for parking. Within the last decade, more research and development work has gone into making the systems smarter with added features; one was smart parking search. OSPI services were introduced into the industry around 2015. BMW Group[1] was one of the pioneers of intelligent parking solutions having started their parking research in 2011. In addition to being beneficial for car drivers, smart systems such as OSPI have benefits for the city as a whole for it improves city mobility conditions when paired with sustainable urban mobility policies. One of the benefits of smarter navigation with OSPI is the mitigation of impacts brought by cruising for a parking space [5]–[8]. The goal is to reduce the amount of cruising vehicles, which constitutes around 30% of the congestion in urban areas [9], [10]. By doing so, there is a direct reduction in air pollution, noise pollution, and unnecessary delays [3]. Furthermore, pre-departure information from OSPI about parking situation at destination helps drivers decide the mode of transport they could take and may even choose to leave their cars behind [1]. Nevertheless, only proper measurement of the quality of such information systems determines the benefits gained in a transport network as discussed by Gomari et al. in [1].

Despite the current benefits of OSPI, the level-of-service, quality and reliability remains an on-going challenge within the area of smarter navigation [4], [11]. One reason is the reliance on manual ground truth data, which presents a trade-off between accuracy and scalability. To address this problem, the goal is to assess the quality with a scalable data collection strategy that uses smart systems and less on-site surveyors, to validate the accuracy, and thereby enhance the information system. This is even more important as the industry shifts towards a future of connected, cooperative, and automated vehicles [12]. Connected and cooperative intelligent transport systems (C-ITS), such as OSPI, have the potential to efficiently and better distribute vehicles within a transport network and achieve a traffic state that is closer to the system optimum. As this paradigm shift takes place, smarter navigation systems must be able to improve and maintain the quality of information that will be shared in cooperative transportation systems. Ultimately, these cooperative and adaptive systems together with proactive innovation and sustainable governance will minimise impacts of traffic congestion and ensure a smart and sustainable urban mobility in cities.

1.2 Scope & Objectives

Given the state-of-the-art, this dissertation intends to address the gap in research for a scalable quality assessment method specifically tailored to mobility-related information systems, and particularly, to tackle the use case of OSPI and its further development.

[1] https://www.press.bmwgroup.com/global/article/detail/T0220542EN/bmw-connecteddrive-develops-intelligent-parking-search-solutions-the-next-step-in-connected-navigation-%E2%80%93-on-street-parking-prediction?language=en, accessed 20 February 2022

The focus is on quality assessment of OSPI parking prediction models; parking search is not addressed and is out of scope. That said, this doctoral research aims to:

- *Further* the understanding of on-street parking behaviour in cities by using large datasets of parking events;
- *Develop* a prioritization-based quality assessment method to address scalability issues with current manual quality assessment strategies; and
- *Improve and enhance* on-street parking prediction models data-driven features extracted from parking events data and enhanced parking-related features.

To address the stated objectives, this publication-based dissertation entails three peer-reviewed studies [1], [2], [4] that all together provide the process from understanding on-street parking behaviour, developing parking prediction models up until the development of a quality assessment that serves the user needs of the OSPI system.

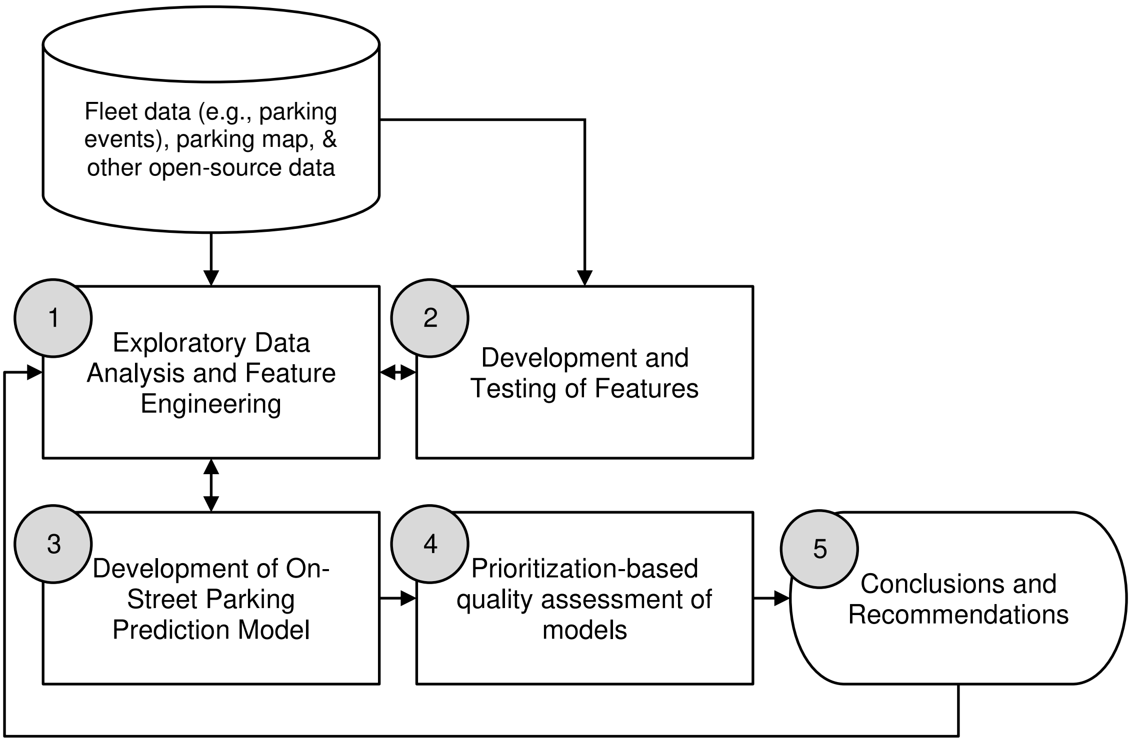


Figure 1. Main research components of the dissertation (source: original figure)

Figure 1 presents the interconnections of the five main contents in this dissertation. As with any data-driven work, there is an input, process, and then output. In this case, as the input data, fleet data are predominantly used. To supplement the inference in all the studies, data is also taken from parking maps and other urban context-relevant data, such as weather information. The input data are used for exploratory analysis to uncover patterns not initially known through unsupervised machine learning (e.g., clustering), and the data are also used to extract features relevant for parking prediction model features. After discovering and engineering new features from the available datasets, the next step is the development of a parking prediction model to comprehend the model development process and the quality assessment gaps and needs. Thereafter, this then feeds into the development of a prioritization-based quality assessment. The last step is then to derive final conclusions and recommendations after conducting the different experiments and studies.

Some parts that were only summarized in the papers are further elaborated in different sections here. This document starts with an introduction is provided to contextualize the topic and offer a general overview, followed by a detailed analysis of the existing related literature in Chapter 2. The developed quality assessment methodology in this dissertation is summarized into a framework in Chapter 3, where the process and interconnections of the methodology's main components are elaborated and connected with the relevant peer-reviewed papers – based on the three main publications of Gomari et al. [1], [2], [4]. Chapter 4 discusses quality assurance by using parking-events-based data-driven features based on [4]. This chapter expands on the main findings of enhanced parking-related features and the role it plays in further developing prediction models into a self-supervising one, which gets updated based on the continuous ingestion of parking events data. Chapter 5 provides a list of topics that could be pursued in future mobility-related information research. Thereafter, this dissertation ends with a conclusion of the research.

1.3 Contributions

In relation to smarter navigation, this dissertation contributes to the field of urban mobility, intelligent transportation systems, and quality assessment methods for software development. Below is a consolidated list of contribution from the three papers of Gomari et al. [1], [2], [4].

- *The cluster analysis of parking behaviour study* [2]. Machine learning is a field that has gained traction over the last decade simultaneous with high performance computers and cloud systems. An important aspect in this field is the usage of domain knowledge to create meaningful features, and consequently, better models. Within the bigger scope of the entire dissertation, the contribution of the cluster analysis of parking behaviour is the development of a methodology for capturing parking dynamics. In the paper, this is called Temporal Trend of Parking Dynamics (TTPD). The concept of TTPD enabled the research to infer parking behaviour from parking events data, and, led to the discovery of new enhanced data-driven features for parking prediction models, as presented in Gomari et al [4].

- *The prioritization-based quality assessment study* [1]. Current ground truth collection strategies are deemed not scalable attributing to costly methods using manual observations. Thus, as stated in Gomari et al. [1], [4], an automated methodology is sought that can reduce the reliance on manual data collection methods, and thereby, reducing associated costs. In line with that, the main contribution of the novel quality assessment methodology is to fairly assess the true quality of mobility-related quality information system prediction models (and more specifically, OSPI), while directing ground truth collection to important areas. The true quality of an OSPI system, as defined in [1, p. 2] is "assessed by assigning importance weights to areas and time periods based on the chosen fleet volume (e.g., parking events, traffic flows)". This methodology avoids potential misfortunate selections that either give an overly positive or negative evaluation of a service. Instead, the approach directs the weights in evaluation score to ensure that data ground truth collection exists in important areas and accordingly, importance, is adjusted relative to the users of the system.
- *The study on development of a data-driven on-street parking information system using enhanced parking features* [4]. Current state-of-the-art models have not used fleet parking events data extensively in OSPI studies. The main contribution of Gomari et al. [4, p. 1] is the value discovery of vehicle parking events-based features to enhance OSPI prediction models. Essentially, these data-driven features enforce quality assurance since data is continuously collected in the backend. Furthermore, their studies also revealed that simple spatial on-street parking capacity features are more valuable when aggregated on a higher neighbourhood (i.e., quadkey) level than just on a street-level. Ultimately, the development of data-driven prediction models and enhanced parking features replace the reliability on historic parking availability features, as presented in [4]. Also, to keep parking maps up-to-date and on a high quality, a parking behaviour change detection (PBCD) model is proposed to trigger alerts in the presence of mid or long-term obstructions to the parking situation in certain areas in a city.

2 Literature Review

This chapter discusses a synthesized and compact literature review. The gaps in literature are identified and reflected upon in relation to the research conducted within this dissertation. This section addresses four main literature areas starting with comparison of the available sources for this dissertation, parking prediction models, on-street parking behaviour change, and prioritization-based quality assessment.

2.1 Comparison of Different Available Data Sources for Parking

The first step after identifying the issue with a scalable data collection strategy was to understand available data sources. A few data sources were identified that could potentially be useful and tested upon. The main attribute looked at was scalability. These included the three readily available data sources including: manual ground truth observations, ultrasonic sensors data [13], and parking events [1]–[4]. Other identified sources that were initially considered included: floating car data [9], satellite [14]–[16], and camera [17]. The readily available data were favourable as these were already widespread and the accuracy could be directly measured. Meanwhile, the latter three sources were not readily available, and an extensive plan was necessary to extract accurate and widespread data, hence, scalability was a concern. Thus, only the first three available data sources were qualitatively analysed by Gomari et al. [18] considering the following criteria: (1) spatial and temporal coverage, (2) technical complexity, (3) costs, and (4) potential for scalability. The potential issues that could arise from using the data sources were also analysed in the comparison. The definition as described by Gomari et al. [18] of each criterion are as follows:

- *Spatial and temporal coverage:* check for the area coverage and time distribution within a city. The analysis output on this criterion contributes to further analysis on correlation of parking spots and POIs considering the density spread of available validation data.
- *Technical complexity:* each data source was evaluated based on the difficulty in processing for it to become useful.
- *Data costs:* the costs associated with data sources were evaluated qualitatively relative to data availability within BMW.

- *The potential for scalability*: the scalability was assessed regarding by weighing the importance of the above aspects relative to the data gathering method and primarily the easy to scale for dynamic data (i.e. detecting open spots).
- *Potential issues:* describing the potential errors and shortcoming of data source.

A summary of the qualitative comparison of the different dynamic data sources are shown in Table 1. All data sources that were rated as high in technical complexity were left out mainly because of the difficulties to acquire them in a fast and widespread manner on top of processing and validating.

Table 1. Qualitative Comparison of Dynamic Data Sources based on Gomari et al. [18]

Data Source	Spatial & temporal coverage	Technical complexity	Costs	Scalability of data	Other potential issues
Manual ground truth collection	Low	Low	High	Low	Human error
Ultrasonic sensors	High	Medium	Low	Medium	Detection issues
Park events	Medium	Medium	Low	High	Does not capture streets that are full
Extended floating car data	High	High	Low	Medium	Volume versus value
Camera	Low	High	Medium	High	Low spatio-temporal coverage
Satellite imagery	Low	High	Medium	Low	Distorted images, necessary corrections, low frequency of images

1. *Manual ground truth (GT)* is the most used method to collect data. Its biggest downfall is the low potentially for scalability, although it is one of the most accurate data for validation. Costs associated with manual collection also tend to be high, thus, it is not a great candidate for scaling.
2. *Ultrasonic sensor (USS)* data in contrast to manual collection has a high spatio-temporal coverage and is low cost in terms of collection; USS data are automatically sent back as geolocation coordinates to the backend. The main issue USS data faces based on analysis is the difficulty to extract accurate detections fast. There are still uncertainties as to the detections. More filtering is required to narrow down the sensor detections to parking spots, and thereby, allow scaling.
3. *Parking events (PEs)* are rated to have medium-level spatio-temporal coverage. This is mainly because only the geolocation coordinates at the final stop or start are sent back. PEs makes up for not having a high coverage by sending back accurate information in comparison to USS. Likewise, as USS, it is also cheap in costs. Since the PEs dataset if already trimmed down to send locations about high potential open spots, it was assessed as the most scalable among the resources available. The biggest issue with PEs is that streets that are full cannot be captured since no parking events will occur there. Knowing this disadvantage, the dataset cannot be used for validation, but can be used as a feature in prediction models instead. The PEs used in this dissertation are further described in Section 3.2.
4. *Extended floating car data (XFCD)* is a source that has been used for many urban mobility studies including parking [19], [20]. XFCD is a collection of geographic positioning system (GPS) coordinates of a car's drive trajectory. It is highly used to extract traffic state estimation [21], as speed can be captured easily. Despite the widespread collection of such data at car companies, just purely based on GPS points from XFCD, it is complex to extract valuable insights for parking prediction quality studies. Two studies [20], [22] proposed to use floating car videos, i.e., videos captured by a moving car in a study area. Although accurate for ground truth collection, these were only used for particular studies to calibrate parking ticket vending machine data and was not necessary the main data source. Since video data is still not automatically collected by car companies, the scalability of relying on it is still not feasible.
5. *Camera video data* as abovementioned objectively speaking can replace manual counts made by humans. However, due to its low spatio-temporal coverage, it is not still scalable until the data is collected and processed automatically. Associated costs can be expensive as well due to the large volume of data

collection from videos. Although this might not be the case anymore in the years to come.

6. *Satellite imagery* has been used in some studies to detect parking spaces [14]. Extracting parking-related information can be quite complex due to distortion and lighting issues. There is also low spatio-temporal coverage, hence, it is not the best for parking prediction quality studies. Given these inherent issues, its scalability is low considering that it is not easy to get imagery from the relevant sources.

All the compared data sources present their own advantages and disadvantages. Based on the available resources at the early stages of data acquisition of this dissertation, it was decided that the PEs will give the most fruitful results, as the other sources did not yield usable results particularly for parking information quality assessment.

2.2 Parking Prediction Models: A Compact Literature Review

Over the last six years since OSPI services became popular, many parking prediction model studies have been conducted. The prediction model studies on a high-level can be categorised into two information systems: off-street and on-street parking. Gomari et al. [3, p. 1] describes off-street parking information as guidance systems leading vehicles to park away from the street into open-air parking lots, indoor parking garages, or multi-storey parking; the parking prediction models for off-street parking estimate the occupancy of these parking lots [8], [23]. On-street parking information on the other hand is described as a parking guidance system that direct drivers to probable open spots that are on-street kerbside parking spaces [3, p. 1]. Normally, a prediction model at its deployment stage consists of five main components as shown in the below (see Figure 2). The elaborated and complex prediction model development is discussed in Chapter 4. In all models, regardless of it being off- or on-street parking, the first component is the training and testing parking data. The next part is generation of features that are fed into a model algorithm, which then gives out predictions. The last step is the evaluation and normally when quality is assessed using a metric.

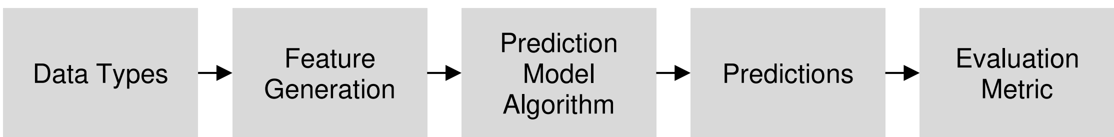


Figure 2. Five main components of a production-ready prediction model (source: original figure)

Same as other fields, in parking prediction studies, the rise of big data has created a shift to data-driven solutions. Researchers have used a variety of data sources. Figure 3 displays an overview of the various data types, Figure 4 illustrates the different models used, and Figure 5 shows the quality metrics (i.e., not necessarily method) used for prediction models.

This subsection provides a compact literature review regarding parking prediction models and the reflection on the current state of literature. In each subsection a summary of the studies is provided followed by more details from selected studies. The summary sections provide the popularity percentage based on the number of citations among the literature reviewed about the data types used in parking prediction (see Figure 3), the algorithm employed (see Figure 4), and the different metrics used (see Figure 5).

2.2.1 Data types used for parking prediction

The general trend among the reviewed literature is that off-street parking studies mostly rely on historic data, while on-street parking prediction requires more than just historic data, as movements and availabilities are more difficult to capture. On-street parking prediction requires models that can generalise based on the spatio-temporal context. In line with that there are two main data categories used: (1) parking ground truth data sources and (2) other supplementary data types that help a mode in prediction.

There was a combined total of 62 mentions of different data types in all the reviewed parking-related studies – this also includes studies that are not related to parking prediction. As presented in Figure 3, within the top 3 most popular data types used according to literature are parking sensors at approximately 24.2% (i.e., mentioned in 15 studies), carpark data at 12.9% (i.e., mentioned in 8 studies), and crowd-sensing at 9.7% (i.e., mentioned in 6 studies); this covers 46.8% of all the data that has been employed in different researches. Despite the popularity of parking sensors in studies, which is mostly based on openly available data, Gomari et al. [4, p. 1] mentions that this data source is particularly associated with high installation costs as most are in-ground sensors. Hence, transferability and scalability are a concern.

Parking sensors are used in [23]–[29], more specifically, real-time ground sensors are used in [26], [27], and some researchers emphasize that parking sensors are not enough especially for on-street parking prediction, thus, the data needs to be contextually enriched [30]–[32]. Another study that relied on on-street parking occupancy system [33] also falls in the category of parking sensors. Furthermore, a research work used parking capacity and parking events from parking bays [34].

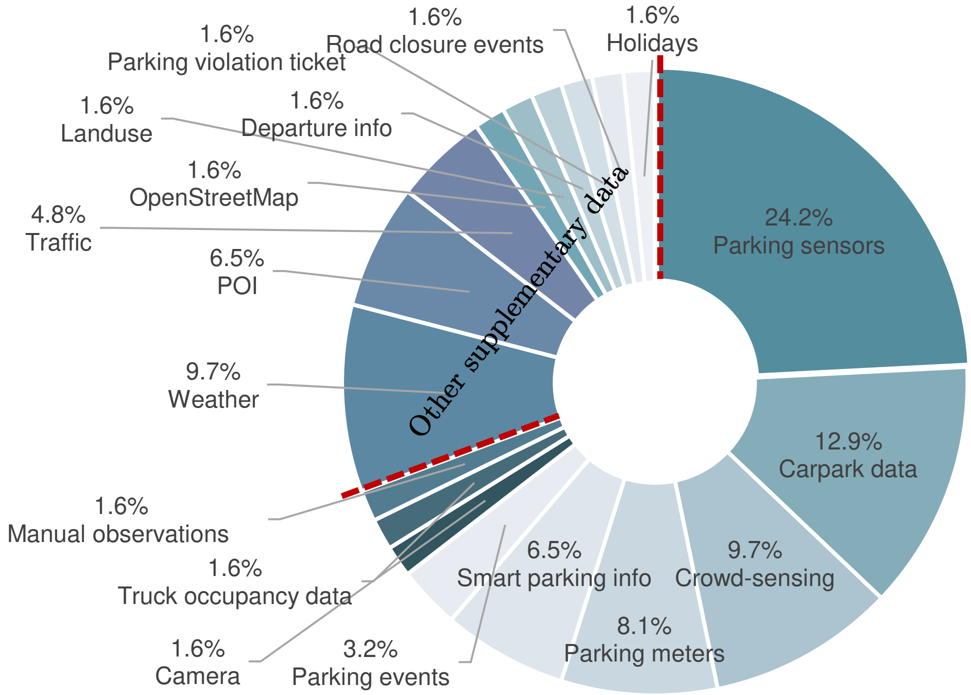


Figure 3. Parking ground truth data sources and other supplementary data types used in parking studies (source: original figure)

Top 2 in the list as mentioned is primarily off-street carpark data [35]. Other forms of car park data used in literature are carpark transactions or payments [36], parking sensors within parking lots [37], and parking occupancy information of parking lots [11], [38], [39]. Another form of carpark data can be counted when a car passes a gate – the study of [40] used such parking records.

Top 3 in the list is crowd-sensing [41]–[43], wherein mobile payments [6], [41], [44] also fall into. These data sources primarily focus on extracting value from large number of devices or applications that are deployed for human input. Inference also plays role here. Typically, if data from crowd-sensing is used, it is combined with other forms of data to strengthen the inference. The studies by Gomari et al. [1], [2], [4], where parking events was used, can be treated as crowd-sensing as well. This falls in the category of crowd data from fleet of vehicles.

Another 22.6% of the data sources are parking ground truth-related data sources. A number of researches used smart parking meters [24], [45]–[48] as the main source for their prediction models, wherein a study like Gomari et al. [2] extracted parking behaviour information using meter transaction [46]. Four studies utilised ground truth smart parking occupancy information [32], [45], [49], [50], two studies used vehicles parking events [51], [52], and one study, among those reviewed in published research, used camera data mounted on a vehicle [53]. The most common data used in practice, as opposed to in research, is manual observations as used in the research of [54]. Majority of the parking studies either were for on-street or off-street parking for private vehicles, however, there was one study that focused on estimating truck parking prediction using historic truck occupancy data [55].

Within the top 3 other supplementary data sources that are used to get more contextual insights are weather comprising 9.7%, points of interest (POIs) with 6.5%, and traffic data at 4.8%; in total they entail 21%.

Weather data [24], [29], [35], [36], [40], [56] that has mostly been used in the studies are temperature and rainfall. In some cases, wind is also used. Although extreme weather may have an impact on parking situation, it is not clear how helpful normal weather data is for day-to-day parking prediction. As demonstrated in Gomari et al. [4], weather data, specifically, rainfall and temperature were insignificant for the parking prediction models built.

In many mobility-related studies, POIs are a great dataset to infer the probability of the activities undertaken within a certain region. In parking studies, only a handful have so far used POIs [37], [45], [57] to improve models. Within the duration of this dissertation, OpenStreetMap's POIs were experimented on in relation with parking behaviour analysis and prediction models. However, it was discovered that open data POIs in general tend to be outdated, while the updated ones are skewed towards restaurants. This problem was even more difficult to solve, as the COVID-19 pandemic heavily changed the opening hours of many POIs.

Another important secondary data source used in studies are traffic-related data including traffic cameras [36], traffic data services [38], and traffic value [29] of each street from the parking information database. Although, one might think traffic data can help parking information, this is not always the case. A study [58] claims that traffic only has secondary influence on parking information. This is consistent with the assumption in this dissertation that OSPI is difficult to predict as contextual data on a local level is needed. Traffic data gives extra information about congestion on a street; however, this does not highly correlate with chances of finding open parking spots.

The rest that makes up 9.6% of the total data sources used in parking studies are composed of data about landuse [45], departure information [40], OpenStreetMap highway data [45], parking violation ticket [57], road closure events [29], and holidays [40].

Overall, different data sources have been used to infer parking behaviour or predict parking availability. Often data are only available in specific areas, hence, it is difficult to transfer results elsewhere from the conducted studies. Also, within the studies conducted, there is no mention about how other data sources can be used to cross validate the models that are based from the primary data source. This happens since normally researches only have access to one source of parking ground truth data. This gap is addressed in the studies of Gomari et al. [1], [2], [4] as part of this dissertation.

2.2.2 Prediction model algorithms

There are many different types of models that are used in researches as presented in this subsection, but there lacks evidence about the models mostly used in practice. Usually, there is a gap between the two since the goals are different. In research, majority aim for optimization up until the tiny fractions, but in practice, runtime and reliability are more important, so long as performance and quality are not compromised as much. Nonetheless, in this subsection, the focus is on state-of-the-art in researches conducted. The frequency of model usages in studies does not necessary mean it is better. Below Figure 4 presented the breakdown of the models used.

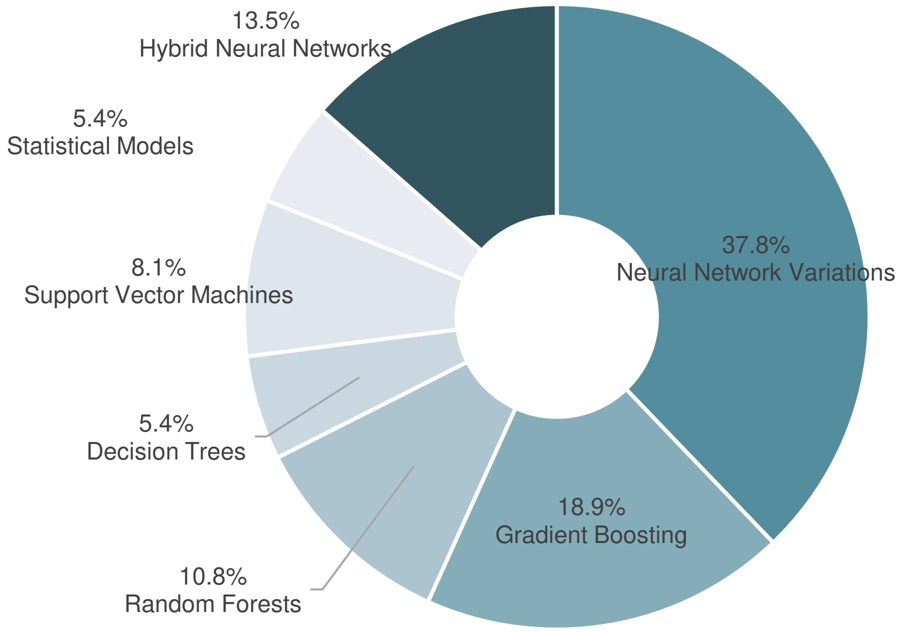


Figure 4. The various machine learning prediction algorithms used for parking prediction (source: original figure)

There was a combined total of 37 mentions of different prediction models in all the reviewed parking prediction studies. The focus in many of these studies were creation and optimization of algorithms rather than experimental discovery of the best input features. The most frequently used models within the top 3 in researches are neural network models accounting for 37.8% (i.e., used in 14 different studies) of all, gradient boosting algorithms are used in 18.9% of the studies (i.e., used in 7 different studies), and random forests are used in 10.8% (i.e., used in 4 different studies). Hybrid and ensemble models are becoming a more promising method as 13.5% of the studies created their own combination of models. Other models that cover 19% of the studies have used decision trees, support vector machines, and statistical models.

The most common time of neural network (NN) [36], [59] model used in researches is Long Short-Term Memory (LSTM) [35], [39], [40], [50], [55]. LSTM is a variant of Recurrent Neural Network (RNN) [35] normally used for time-series prediction problems, although it not necessarily the best model, considering the processing time. Specifically, according to [40, p. 267] LSTM "[...] is characterized by adding valve nodes of each layer in addition to the RNN structure to achieve long-term and short-term memory functions, so it performs multi-step predictions of multivariable time series" – this enables the model to remember localised patterns also in the longer term. Other types of NN used in research are Wavelet Neural Network [48], Convolutional Neural Network [36], Hierarchical Recurrent Graph Neural Network (GNN) [37], and Artificial Neural Network (Feedforward Neural Network FFNN) [45], as also used in Google's research team implemented as a single layer regression and feed forward deep neural network [60] for estimating difficult of parking using mainly google maps travel data.

More commonly machine learning algorithms used in practice are tree-based models like Gradient Boosting Regression Trees. With the introduction of Xgboost [61] in the industry, many companies and products trusted the algorithm as it provides reasonable default parameters and the runtime of the model even with big data is fast. This means, teams do not necessarily need to tweak many parameters before initial launch. Nonetheless, hyperparameter optimisation of Xgboost is still vital in production machine learning models. Parking researches have used the following: Gradient Boosting Decision Trees (GBDT) [48] if the model is framed for a classification problem, Gradient Boosting Regression Trees (GBRT) [11], [29], [62] when dealing with continuous probability models, Xgboost [11], [38], lightGBM [11], and CatBoost [30]. These are all variation of different gradient boosting algorithms. According to literature, Xgboost is the most performant among all the variation. Although, there are new variations of Xgboost that are currently being developed for more specific applications.

The third most popular parking prediction model are Random Forests [11], [30], [36], [45]. Random Forests have been around since 2001 [63]. It is basically a better version of Decision Trees [49], [57], which are also used in parking studies. The biggest advantage of Random Forests and its fame in usage is its simplicity in implementation and that there are only a few parameters to adjust since the model relies on a forest of decision trees, it is intrinsically already finding patterns among different combinations of features in the ensemble models built.

Another approach to parking prediction model studies is using hybrid models. Often this entails a combination of either multiple NN or NN with Gradient Boosting models. A study [34] developed a hybrid Graph convolutional network with gated linear unit. Another combined Xgboost with an LSTM model [55], while acknowledging that Autoregressive integrated moving average (ARIMA) [64], Regression Trees, and Neural Networks are the most common models used in practice. Another study created a hybrid

CNN and LSTM model [50]. A step higher than that, a group of researchers combined multiple graph CNN (MGCNN) with LSTM [31]. It is not surprising that hybrid models are becoming more popular. There is definitely something a certain model is not able to capture as also proven a comparative study [59] that analyses different parking prediction models and ensemble learning models that combined the different individual models. A research [65] conducted a comparative study of different models and combined and tested against each other models comprising of linear regression, support vector machine, neural network and ARIMA. Despite the advantages of ensemble models in better accuracies, a couple of disadvantages of such models are the slower processing time and also the inability to understand and interpret the reasons behind improvement.

Two other approaches that are not that common in research are Support Vector Regression (SVR) [32], [57], [66] and statistical or mathematical approaches. A study for example [25] used availability mean with variance, normally distributed availability, normally distributed availability variation, and non-homogeneous poisson distributed arrivals and departures. This approach was also able to generalise parking situations that fed into parking availbiltiy.

As shown in literature, neural networks are a popular approach in implementing parking prediction models. However, demanding runtime requirements still make it unattractive for production deployment and fast testing – this may not be the case in the near future as cloud systems are becoming more affordable. And interpretability is still an issue with neural network models. A study [45] even got results proving that Random Forest outperformed Artificial Neural Network for their parking prediction use case. This is proof that even with longer "thinking time", a more complex and sophisticated model is not always the best option. It must also be noted that, implementation of neural networks is also a timely process (e.g. undestanding number of layers needed) in comparison with Random Forests and Xgboost, which can be used with a few hyperparameter tuning. Furthermore, there are methods to extract feature importance that enables interpretability of the models.

Most studies reviewed focus on determining probability based on parking occupancy. This means, the focus is to estimate the percentage fullness of a certain street from 0 (empty/completely vacant) to 1 (full/completely occupied). This approach shows the customer of the information system how many slots have already filled up. Another approach taken by Gomari et al. [1], [4] is to predict whether there is at least one available parking spot or none. The argument is that customer only care about if they can find one open parking spot. It does not matter if a street is 10% full or 90% full, as long as there is one open parking spot. The availability probability approach is much more difficult, but this dissertation focuses on this approach.

2.2.3 Quality assessment metrics

There is a difference between metrics and methods. Metrics solely one its own can reflect a false representation of the quality as argued by Gomari et al. [1]. A quality method as that presented in [1] is a process that ensures the quality of a product is up to the desired level to meet customer satisfaction. According to the best knowledge of Gomari et al. [1], there are no quality methods specifically for mobility-related information systems and particularly, on-street parking information. Thus, this subsection focuses on the quality metrics that have been used in parking studies. Nevertheless, metrics are still part of quality assessment methodologies. The distribution of different quality metrics used in the studies reviewed are illustrated in Figure 5. The formulas for the metrics are not tackled here. For details on the use of each metric in the studies reviewed, refer to the citations directly.

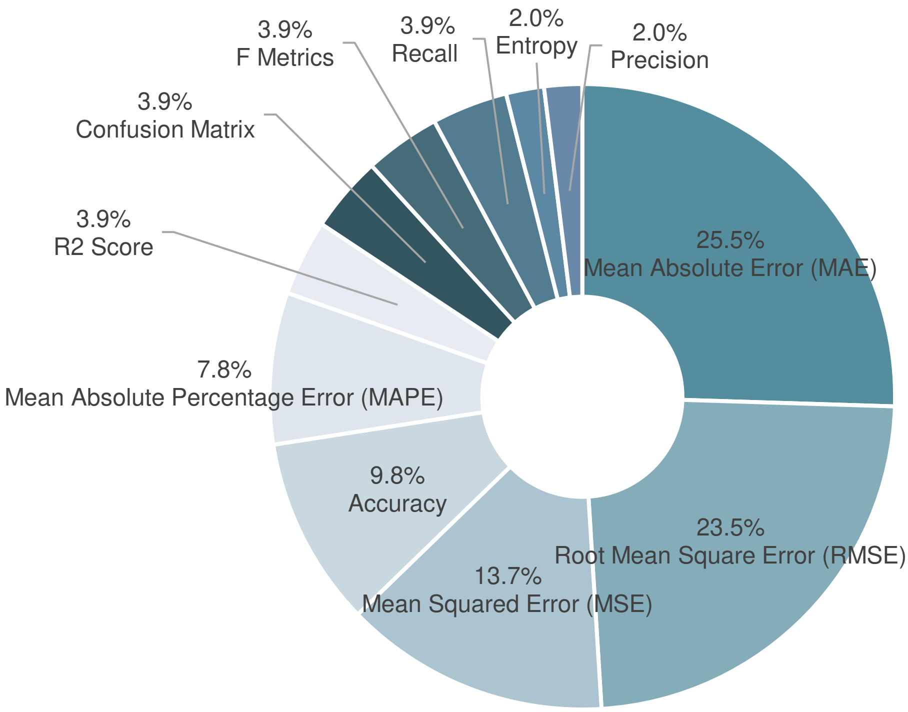


Figure 5. The different metric scores used in studies (source: original figure)

There was a total combined of 51 mentions of the different metrics used in parking-related studies including those that are not related to parking prediction. The top 3 metrics used in literature cover 62.7% of all studies. These metrics are Mean Absolute Error (MAE) found in 25.5% of studies (i.e., used in 13 different studies), Root Mean Square Error (RMSE) in 23.5% (i.e., used in 14 different studies), and Mean Squared

Error (MSE) in 13.7% (i.e., used in 7 different studies). These metrics are commonly all measured together in many studies. Besides the use of MAE in many parking-related studies [11], [24]–[26], [34]–[37], [39], [45], [62], one study even derived a metric from MAE, that is, Mean Absolute Scaled Error [67]. MAE, in general, is a popular metric used between models to determine the absolute difference between observed and predicted values. Shortly after MAE, RMSE is used in research [26], [29], [32], [34]–[37], [40], [55], [57], [68] with one study [28] having special variations of RMSE, Normalized Shannon Entropy of RMSE, and the harmonic mean of the two labelled as F. RMSE is essentially the standard deviation of residual and is measured by taking the square root of MSE. MSE [27], [35], [36], [39], [45], [48], [62] is also equally popular in parking studies and measures the variance of the residuals by measuring the squared difference between the observations and predictions. The application of MSE in studies where the prediction is concerned with probability, hence, having values between 0 and 1 is called Brier Score, which is used in Gomari et al. [1]. The top 3 metrics already cover majority of the studies, and based on that, it is easy to determine which metrics are deemed useful in parking prediction models.

There are 11 other metrics used in the remaining 37.3% of studies including: Accuracy [40], [42], [44], [53], [59], Mean Absolute Percentage Error (MAPE) [24], [26], [34], [39], the coefficient of determination or R2 score [30], [36], [45], Recall [49], [59], F metrics [32], [59], Precision [59], Entropy [32], False Positive Rate (FPR) [25], False Negative Rate (FNR) [25], Receiver Operator Characteristic (ROC) [25] and Confusion Matrix [42].

The review of metrics is to show that, not many studies expound on the quality assessment or evaluation of their models. But there are some authors that explain their reason for choosing a certain metric, for instance, Balmer et al. [45, p. 4] mention that MAE and MSE are good measures of difference in magnitude instead of using a metric that measures relative percentage difference such as in MAPE that will focus predicting small occupancy values, despite high occupancy values being more important. This is exactly the problem Gomari et al. [1] solve by introducing a prioritization-based methodology to adjust any metric used corresponding to when it is most critical for the users to get information. No study besides this dissertation as conducted based on Gomari et al. [1] has given a solution to adjusting the final metric by considering the relative to user importance problem. A few studies [45], [49] do acknowledge that busy hours are the critical hours to have accurate information but did not device or could not use a better method to adjust the scores. The main reason is the lack of data that allows prioritization. Gomari et al. [1] use BMW data of fleet parking events that allows for identifying important areas.

2.2.4 Discussion

A general finding from reviewing the evaluations of different studies shows that off-street parking studies have higher evaluation scores in comparison to on-street parking. The reason for this is the different factors that are influencing on-street parking as comparison to off-street parking. It was mentioned by several studies that it is important to contextually enrich [30]–[32] ground truth data in order to create better OSPI prediction models. There are many different prediction model algorithms (see Figure 4), but in this dissertation, the focus was not on building or optimizing algorithms, but it was to focus on the utilisation of data and the features engineered within prediction models employed in the context of a parking information systems. An important aspect in algorithm of prediction model selection was to allow interpretability. This was possible with tree-based models like Xgboost and Random Forests, and not with neural networks. Besides, an important approach in this dissertation was to utilise domain knowledge by first understanding parking behaviour and methods to cluster the time-series data from parking events as presented in [2]. This is directly related to enriching the models, but also comprehending the process of enhancement as presented in [4].

There is significance in understanding the parking behaviour to influence development and build parking prediction models [2]. Balmer et al. [45] studied the importance of geospatial data for prediction models. Jelen et al. [30] discovered that contextual information makes their predictions better. Gomari et al. [4] describe that predictions are made based on the features given to them. These features are extracted from data that is fed into them as shown in Figure 3. The better a researcher understands the behavioural nature of parking, the better features one can come up with to design their behaviour in a prediction model. The main step to understanding the different aspects of OSPI prediction models and their quality assessment was researching on parking behaviour. The aspect of parking behaviour sets up the framework that is built for developing a quality assessment methodology focusing on the OSPI use case.

More data fed into a prediction model does not necessarily translate to better and faster solutions. As presented in Section 2.1, each data source has its own benefits and drawbacks. The data used for any application must be selected with precaution. There is a need to have sound domain knowledge [69] to develop better products. Domain expertise likewise help engineer better features to deal with big data for supervised machine learning [70]. Although feature engineering is a manual and time-consuming task, it can make a substantial difference in the performance of supervised machine learning models, such as those used for on-street parking prediction. Gomari et al. [4], for instance, after doing an exploratory analysis as part of a cluster analysis study on parking behaviour [2] discovered that developing new features from "...street parking capacity features play a major role in the performance of the models" [4, p. 10].

Fundamentally, the primary difference between state-of-the-art OSPI models as described in is how the data available are used for training, testing, and validating the models [1], [59]. As Gomari et al. [4, p. 1] mention: “the differences in input data play a major role in the reliability and quality” of a system. Thereby, this is the step that must be given the most attention, and not aiming to build prediction models that are difficult to interpret.

2.3 On-Street Parking Behaviour Change Detection (PBCD)

Change or anomaly detection studies are quite common in the field of machine learning. It is more regularly used for anomalies noticed in continuously ingested data at the server side to detect outages or to account for missing data. The process is usually generalized in two steps: the prediction step with machine learning algorithms and the second step where anomaly detection rules are set to identify outliers in the data. Gomari et al. claim that there currently [4, p. 2] “no known studies that specifically use parking events to determined potential changes in parking behaviour associated with longer term static changes like in rules and restrictions, constructions, or infrastructural changes”. A study [41] used sensor data to detect unusual patterns and infer possible disturbances in parking location sensors – as mentioned, related to missing data or outage. Another research [32] used park-out events for anomaly detection of legal and illegal parking spots in comparison with their map. But neither of the studies conduct PBCD to improve parking prediction model and its quality

A study by Shipmon et al. [71] at Google investigated such anomalies using streams of traffic data. Their research used machine learning to create regression models and predict the pattern in data, and as a second step they set anomaly detection rules to identify anomalous changes. In the case of parking studies there are only a few researches that have focused on leveraging data to detect anomalous behaviour. Bhattacharyya et al. [66] looked at real-time anomalies detected from sensor data in various garages across Santa Monica, USA. These were controlled environments and every car coming in and out was counted. There were other studies that detected unusual patterns in parking behaviour using parking sensor data [66], [72]. Domakuntla [51] as part of this dissertation's research umbrella studied in detail the potential of using parking events as a data source to detect static and dynamic parking behaviour changes. Gomari et al. [4] leveraged parking events data for static parking behaviour changes as an added feature for a data-driven OSPI model.

A gap in parking prediction data-driven models is internalising change detection to keep OSPI services up-to-date. A service provider [73] mentions that maps are only updated every quarter. This is because of due diligence as it is a costly process. Many parking prediction models use features that best generalize a population. This means, sudden changes are difficult to detect unless a large volume of data is collected rapidly

[51]. That entails having a large fleet of vehicles sending signals from everywhere in a city or study area. To address this problem, a solution is proposed by Gomari et al. [2] that focuses on aggregation of data to increase volume of data from surrounding areas and time periods to identify spikes that are irregular compared to recurring patterns. Particularly, Gomari et al. implements this methodology in [4] using parking events as the fleet data. This solution then addresses the lack of such features for automated updates, as already existing in some mapping services for traffic information and routing.

The study of [51] looks further into analysing the potential usage of parking events in developing a dynamic PBCD model. The study looks into adding real-time features such as traffic congestion information obtained from HERE maps and rainfall data. The PBCD is modelled as a time-series problem and partially using clustering as done in [74]. Once more data can be gathered or acquired from multiple fleets of different car OEMs, it would be possible to build a reliable PBCD model. Essentially, the gap is the lack of sufficient data. This can also be addressed once camera data becomes more prominent in the development and improvement of OSPI services.

2.4 Prioritization-Based Quality Assessment

As reviewed in this literature review section, there is an apparent gap in modern quality assessment methods other than a simple metric. Balmer et al. [45] was among the few studies that described the rationale behind using the metric they employed in their study and also acknowledge there is an issue with properly assessing the quality of a prediction systems. However, they only go as far as mentioning it is a metric issue and not a methodological concern. There are no existing methodologies that specifically look into a prioritization-based approach to address the quality assurance concerns of mobility-related information systems.

Even in modern machine learning practices, the rule of thumb is to use a metric for objective-based optimization. Hence, a model will look to minimise the overall losses based on its available hyperparameters and model architecture. This approach fails to recognise if the optimization is hurting a model and in the case of an OSPI system, whether it compromises having an average model in a difficult area with a bad one, in order to have an excellent model in an easier area to predict. Thus, the importance of user usage is ignored. This gap is completely addressed in the study of Gomari et al. [1], where they propose to use a prioritization-based subsampling approach to adjust scores relative to frequency of parking in a certain area and time period.

3 Methodological Framework of a Prioritization-Based Quality Assessment for Mobility-Related Information Services

This dissertation introduces a new methodological framework for quality assessment of mobility-related information services, and particularly, on-street parking information (OSPI). The framework is based on a prioritization-based approach that ensures models are assessed based on the frequency of parking in certain locations and time periods. This translates to importance for the users of the systems, and thus, the methodology adjusts scores highlighting error prone and subsequently, low-performing areas. This is further discussed in the subsections in this chapter. This chapter also tackles the development process of the framework and the connection of the different elements that are part of this dissertation. Figure 6 shows the mind map and the interconnections of the different topics tackled within the duration of the research period.

The mind map shown in Figure 6 illustrates the major processes and the connection of topics among the three main papers [1], [2], [4] as part of this dissertation. Arrows in the image represent relationships. Links that have no arrows direct to subtopics of its mother topic. The three main studies are represented in three branches stemming from the overarching topic of this dissertation: "Quality Assessment Methodology for Mobility-Related Information Systems". These three are: (1) Exploratory Data Analysis (see Section 3.3), (2) Prioritization-Based Subsampling Quality Assessment (see Section 3.4), and (3) Parking Events-Based Data-Driven OSPI Systems (see Section 4). On the upper portion of Figure 6, to the left of the overarching topic is the database, which was the basis for all the work done in the main topics. In this figure, it is clearly demonstrated how each discovery is linked to the processes in the other studies. For example, within the exploratory data analysis, cluster analysis of parking behaviour was conducted, where two vital components were discovered: Temporal Trend of Parking Dynamics (TTPD) and the quadkey concept. These two as shown in the figure are linked with the other two topics.

The green boxes in Figure 6 represent floating topics. One, as already mentioned, is the database of this dissertation. The other three are main outputs that were achieved within this research. The end point and target of this study as shown in the mind map is the topic on lessening the reliability on manually collected ground truth data.

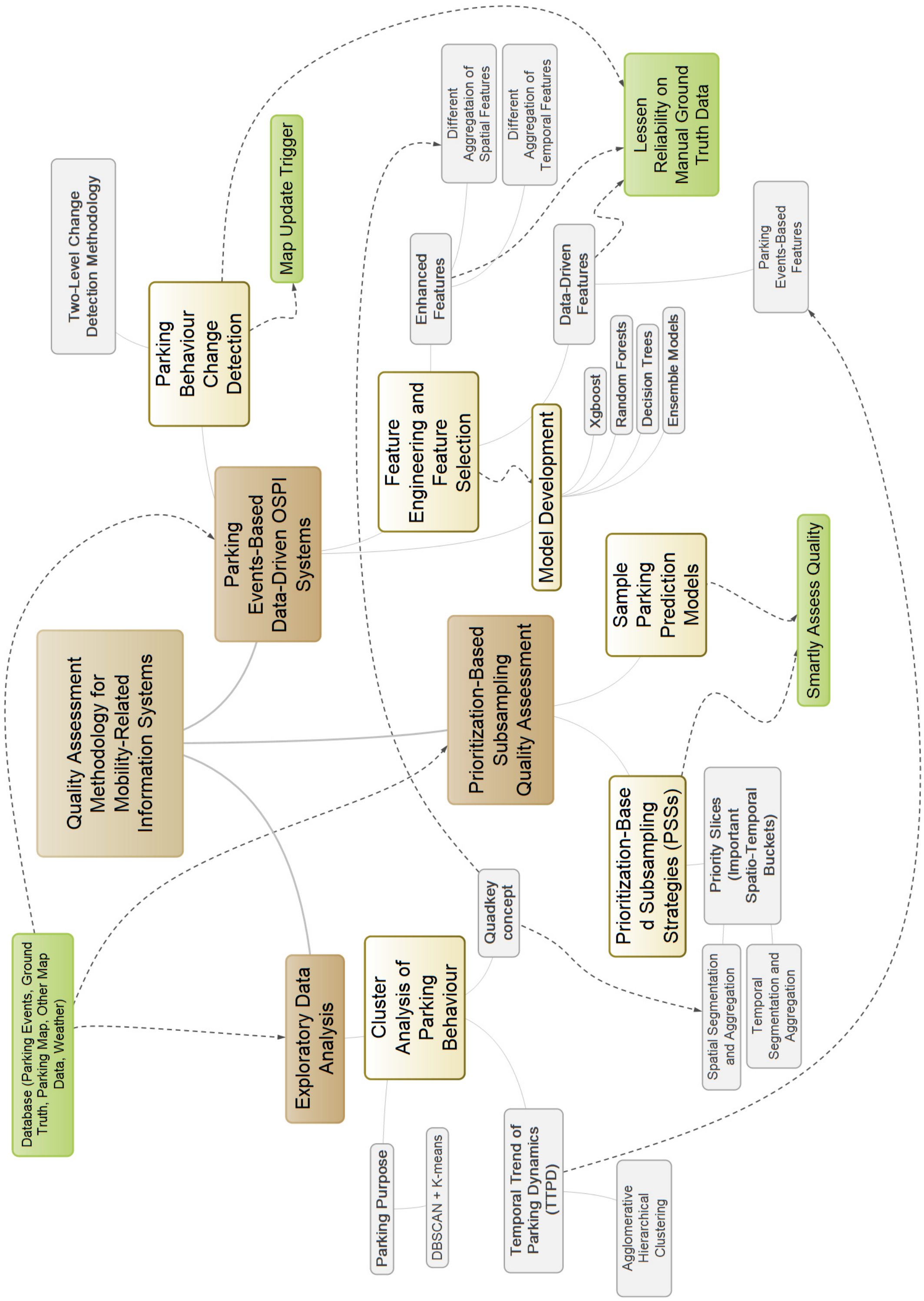


Figure 6. Methodology mind map showing interconnection of all components of the framework (source: original figure)

3.1 The significance and usage of the developed quality assessment

3.1.1 Benefits for product development

Software product development is complex. Not all software products that are developed turn out to be useful. As mentioned by the Lean Startup Co. [75], there are steps to achieving product-market fit; this means developing a product that fits the needs of potential customers and solves their problems. The more inconveniences resolved, the more attractive the product will become. This does not necessarily mean the metric measured ensures seamless product quality experience. The company Heap [76] mentions that "[...] many teams over-rotate on product quality, while overlooking the pool of people willing to buy it". This is the issue specifically Gomari et al. [1] are addressing. They [1, p. 2] address that "despite advances in artificial intelligence, [...] there is still potential to attract more users to increase benefits on a system level". The study highlights the importance of customer-centric quality assessment. Primarily, focusing on improving the value for the user of the information rather than ensuring product interface quality.

3.1.2 Benefits for future mobility-related products

Mobility-related information systems or products will continue to evolve. Society has gone from using compasses and now to built-in software solutions in smartphones – placing convenient navigation on the tip of our fingers. Looking ahead, enterprises, cities, and research institutes have become more proactive than reactive. The impending transition to connected and automated driving [12] will pose many quality assurance challenges – especially, in cooperative scenarios between vehicles. The safety concerns and the value of a product will rely on the reliability and accuracy of the relevant systems for its users. Figure 7 shows where the quality assessment comes into play in the future of urban mobility on a higher level. The developed methodologies in this dissertation can be used to ensure that quality is evaluated correctly. The accurate assessment of systems can instigate quality improvement and thereby, propagate the benefits of the mobility-related information systems to connected and cooperative systems, which ensure gains at the system level of an urban transport network.

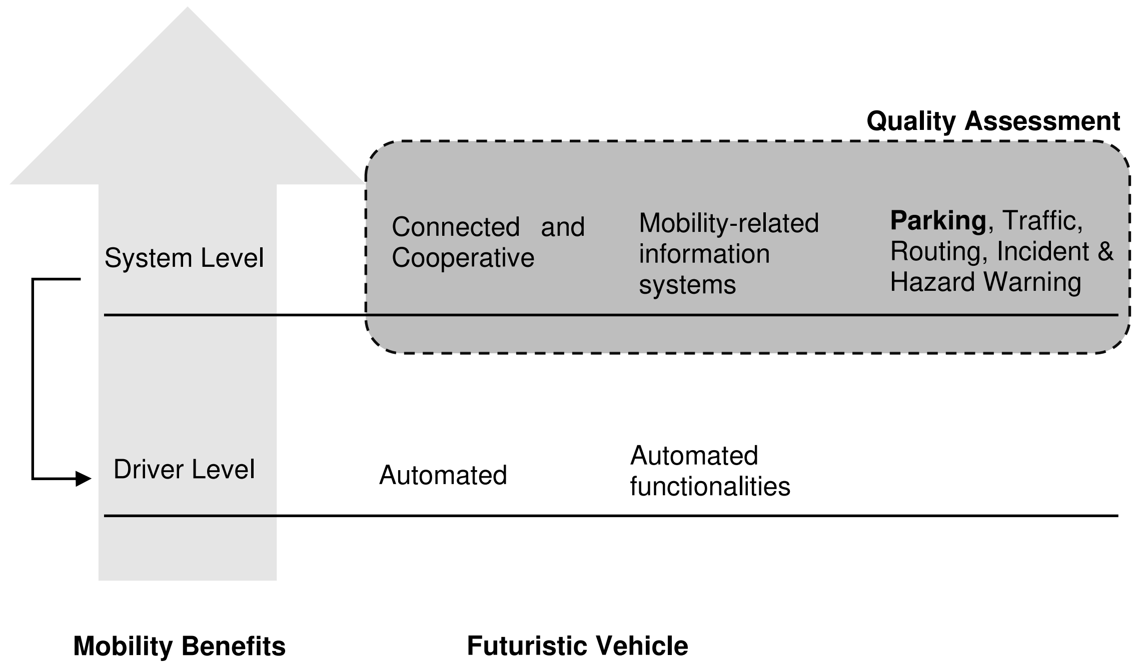


Figure 7. The role of quality assessment in the bigger picture (source: original figure)

3.2 Description of data used and data processing

The data used in this research are categorised into four groups: (1) the Munich study area polygon and its parking transport network, (2) parking events as the main data source, (3) ground truth observations as the training and validation data for prediction models, and (4) other supplementary data. The acquisition and usage details are described below:

3.2.1 Geometric data: Study area polygon, transport network, & quadkeys

Prior to anything else, Munich, Germany was defined as the study area for all experiments. The main reason for this was the immediate usability of the data in 2019; today, in 2022, there are more cities that could be studied. The Munich polygon was taken from the available OSPI parking map in 2019. Likewise, the parking transport network was also taken from this map. These data were the basis for setting the geometrical boundary conditions of this research. In order to have uniform and discrete segmentation of the study area into smaller ones, Microsoft's [77] Azure maps tile standard, also known as quadkey, was used. This is to enable replication of approach in other real-world applications. Specifically, quadkey zoom levels 14 (2.5km x 2.5km) gave the best inferential results from the studies. This approach was used in all three main studies done by Gomari et al. [1], [2], [4].

3.2.2 Parking events (PEs)

PEs, within the scope of this research, are gathered anytime a car switches on or off its engine – translating to a parked-out or a parked-in event, correspondingly. An exemplary CSV table of PEs are shown below in Table 2. The PEs were the primary data source and were constantly being gathered in BMW's backend. The data were also further processed to only capture events that are within a proximity of a street and have durations longer than 5 minutes, which is typical for pick-up and drop-off. The data used has a bias towards BMW drivers; this is acceptable as the data usage is ultimately to the benefit of the same users. Gomari et al. mentioned [4, p. 3] "as opposed to studies reliant on ground truth [26], [78] that cover only certain parts of a city, this research aims to utilize parking events as floating sensors". These floating sensors were used to estimate parking situations and fit into the prediction models as attributes (see Section 4.1). Having an open data source from all car OEMs would be the ideal scenario to have a one-stop-shop floating sensors solution for better OSPI systems. However, many collaboration steps need to be taken, which are not addressed here.

The best attribute the PEs have that makes the dataset highly valuable is the geohash index [79] (see Table 2) – the encoded latitude and longitude geolocation information. In lieu of unique identification numbers of each event due to privacy, this was used to match pairs of parked-in and parked-out events. The ability to match pairs allowed for the extraction of features such as parking duration, parking dynamics, and parking purpose as described in [2, p. 4]. Elaboration of parking dynamics and parking purpose is discussed in Section 3.3.

Table 2. Exemplary CSV table input of parking events

Geohash	Latitude	Longitude	Date & timestamp	Event type
u281z73dmnfg	48.138393	11.570882	2022-02-22 14:59:17	Parked-in
u281z73dmnfg	48.138393	11.570882	2022-02-22 17:32:10	Parked-out

3.2.3 Ground truth (GT)

GT data refer to manual observations made by on-site surveyors. Each city where an information service operates has a defined geometric service region. Observers were normally deployed at random times to collect data. There were two main criteria for counting: (1) count a street as open, if there is one legally available on-street parking space, otherwise, mark as fully occupied, and (2) only count, if a parking space is at least 5 meters long. There was no post processing done on the GT dataset as this was easily gathered using the INRIX app, where the parking street network could be updated with GT counts.

The processing of GT into model features are further described in the comprehensive model development in Section 4.1.

3.2.4 Supplementary data

Gomari et al. [4, p. 4] briefly mentioned that two other sources were also used in feature engineering of prediction models, which included construction information map from HERE[2] accessed in March 2021 and open weather data from Deutscher Wetterdienst[3] accessed in July 2021.

Particularly, the construction data was used as a validation data source for the parking behaviour change detection [4, p. 11] using processed time-series parking events. The parking behaviour change detection model was designed as supplementary information component for the OSPI system; this is further described in enhanced parking-related features in Section 4.1.

Weather data, specifically, temperature and rainfall, on the other hand was directly incorporated as a feature in the prediction models to see whether there would be any impacts on the predictions. The feature selection step continuously excluded this feature, as it could not capture variances relative to the other existing features. It can inferred then than weather does not play a significant role for on-street parking in the city of Munich. The case can be different in cities with heavier rainfall and more extreme climate.

3.3 Exploratory analysis: cluster analysis of parking behaviour

The development process until a methodology (see Section 3.4) was developed involved two exploratory aspects: (1) selecting and further understanding the primary data source, which was the parking events dataset, (2) developing a sample on-street

[2] https://www.here.com/, accessed in March 2021
[3] www.dwd.de/, accessed in July 2021

parking prediction model (see the next Section 3.4.1) to identify gaps that exist in quality evaluation, and (3) design a quality assessment methodology to address the gaps discovered (see Section 3.4).

After understanding the basics about the available data, exploratory data analysis was conducted [2] to get more insights. Particularly, Gomari et al. [2] developed a methodology to infer parking behaviour from parking events. Since the goal was to understand patterns, an unsupervised learning approach was employed to generalise parking behaviour. Estimating mobility behaviour from geolocation data is common in the mobility field as demonstrated in [80]–[83]. As mentioned by Gomari et al. [2, p. 1], the hypothesis that was tested is whether parking events can give more insights about the parking dynamics in the city. There are different methods to achieve this, among which, clustering was selected. There are various clustering approaches [84] that can be used as done in many parking-related studies as well [2], [32], [47], [72], [85], [86].

Prior to clustering, the parking events data was transformed into a time series format and further processed to capture behaviour. The steps taken in the study of Gomari et al. [2] can be summarized as follows: (1) data time series composition, (2) agglomerative clustering on the temporal trend of parking dynamics, and (3) then inferring generalised parking purpose from the application of a two-stage DBSCAN – K-means clustering on the parking duration. The key elements of Gomari et al. [2] are presented hereafter. For detailed analysis, refer to the paper.

3.3.1 Parking events time series transformation to temporal trend of parking dynamics (TTPD)

Having understood that parking events can be paired, a new parameter was developed, namely, the temporal trend of parking dynamics (TTPD) (see Figure 8). Gomari et al. [2, pp. 3–4] describe this as a parameter that can be used to "[...] estimate the activity of parking happening in each quadkey". TTPD is defined as [2, p. 3] taking the cumulative sum, at each quadkey, of the net parking, which is calculated by the summation of the difference between the average parked-in (PIN) and parked-out (POUT) events aggregated on 15-minute intervals over a 168 week-hours. To reproduce this, the following steps need to be done considering a tabular format (or a dataframe in Python):

1. Divide the study area into quadkeys at zoom level 14. For each quadkey there is one unique time-series and the next steps are applied on the quadkey level.
2. Categorize the parking events data into PIN and POUT.
3. Sum up the PIN and POUT volumes at each 15-minute interval.

4. Create a new attribute named week-hour. This is done since based on time series decomposition [2, Fig. 1], it was determined that the dataset has a recurring weekly pattern.
5. In one week, there are 168 hours. Further subdivide this into 15-minute intervals. This results to 672-time bins or time intervals for a week. For example: 7, 7.25, 7.5, 7.75 week-hour.
6. Since the parking events data has a weekly recurring pattern, the average at each interval over the one-month study period is calculated. The output could be a float number.
7. Now, after taking the average, in the tabulated data, each time interval should only have one value representing the average volume of PIN or POUT at that interval.
8. Starting from 0 week-hour the cumulative sum is then calculated with a 15-minute time interval step.
9. The value can be either positive or negative. As presented in Figure 8 when the time series graph is going up (positive slope), this means the quadkey or neighbourhood is filling up, while when the graph is going down (negative slope), this translates to the emptying of the neighbourhood.

Mathematically, the TTPD is defined as in [2, p. 4]:

$$TTPD_q^T = \sum_{t=0}^{T} PIN_q^T - POUT_q^T \tag{1}$$

where $\mathbf{PIN_q^T}$ & $\mathbf{POUT_q^T}$ are time series vectors containing 15-minute aggregated parking events at each quadkey, q (i.e. $\boldsymbol{P_q^T} = \{P_{qt}^T; q = 1, 2, \ldots, N; t = 00{:}00, 00{:}15, 00{:}30, \ldots, T\}$, where N is number of quadkeys in the study araea, and T corresponds to the length of study time period, defined as one week.

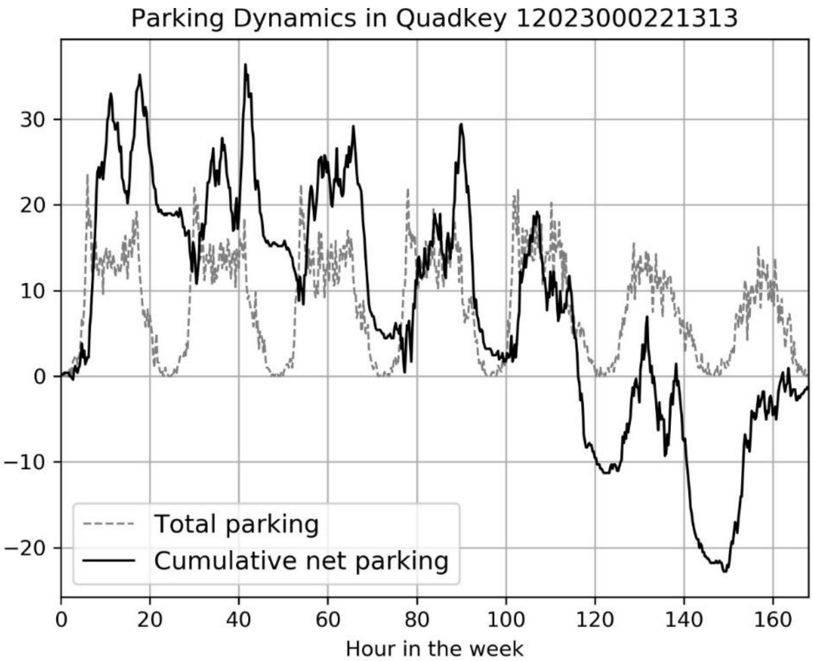


Figure 8. An example of temporal trend of parking dynamics (TTPD) of a commercial neighbourhood (or quadkey), as presented in [2, Fig. 5]

The abovementioned transformation of the parking events data presents the potential of parking events to infer parking behaviour. In simpler terms, the TTPD time-series graph shown in Figure 8 is also called the cumulative net parking in the area. If the value of net parking is positive, it means more cars have parked in that quadkey compared to cars that left that quadkey. Knowing this behaviour and understanding the patterns, clustering can be done to combine similar neighbourhoods that are represented by quadkeys. This transformation is a prerequisite to the clustering presented in the next Section 3.3.2, and also pre-sets the data for application in further feature engineering in Gomari et al. [1], for which the summaries are presented in Sections 3.4.1 and 4.1, respectively.

3.3.2 Clustering of TTPD to infer parking behaviour

The clustering employed in the parking behaviour study of Gomari et al. [2] entails a combination of unsupervised machine learning clustering techniques for parking behaviour inference. Agglomerative Hierarchical Clustering [87] was used for clustering parking dynamics, and density-based spatial clustering of applications with noise (DBSCAN) [88] and K-Means were used for inferring parking purpose. The developed clustering process methodology is illustrated in Figure 9. The implementation is described in [2, pp. 3–7].

For each clustering method, a geometric metric parameter can be selected to quantify the differences between data points that are being clustered. The chosen metric can depend on the use case. For this research, Euclidean distance was used in all instances for similarity or dissimilarity measures.

Agglomerative Hierarchical Clustering was used to see which quadkeys have similar TTPD time series. For this approach, it was not required to enter an initial number of clusters. Using a stepwise approach and illustration through the dendrogram, the number of clusters could be decided afterwards. After clustering 5 meaningful clusters were used later in the analysis that were categorised individually into or a combination of residential, shopping and commercial (i.e., business hours), and dining (i.e., eating).

As for the parking purpose, two parameters were used for clustering the data points: the parking duration in minutes of each PIN and POUT pair, and the PIN time of the pair. Since there were many pair points, DBSCAN was chosen to cluster the points that were truly distinct from one another and have a bigger distance gap. Two big clusters were generated with DBSCAN, separating daytime and night-time parking. To further subdivide these into groups, K-means was utilised to separate short-term, medium-term, long-term, and overnight parking. Details can be found in [2, Fig. 6]. From the total of 16 clusters, parking purpose was then inferred.

In the end, the parking dynamics and parking purpose were cross-tabulated, and this was thoroughly labelled and analysed to understand on-street parking behaviour in a city; in this case, Munich, Germany. Furthermore, the methodology immediately provides an understanding of the spatio-temporal behaviour of on-street parking [2, p. 8].

The results show that using the methodology introduced, the parking behaviour within the city can be obtained using the developed unsupervised learning approach with clustering. The main contribution of this exploratory analysis was that it provided better insights into the limitations and the potentials for the parking prediction development and its usage in quality assessment. For instance, the development TTPD as a parking dynamics parameter is more useful than merely using parking duration instead. TTPD can capture more information, and thereby, has the capability to generalize situations better. This was proven in Gomari et al. [4], and as described in Section 4.1.1. A limitation discovered was that the data was not dense enough to have parking behaviour features on a micro and more local level. Nonetheless, the cluster analysis provided a well-rounded exploration that set the direction moving forward with the next studies.

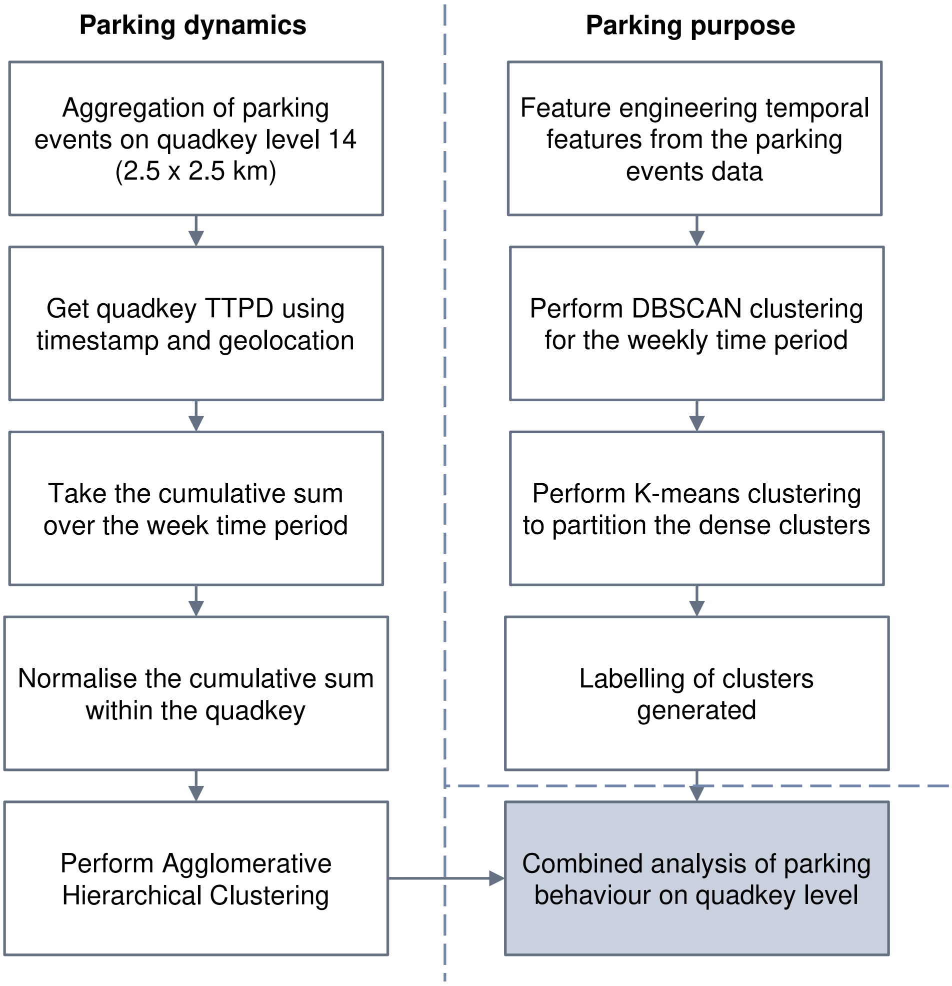


Figure 9. Cluster analysis of parking behaviour methodology, as presented in [2, Fig. 2]

3.4 Prioritization-based quality assessment key components

The exploratory analysis sets up the research on how to utilize the parking events data. Gomari et al. [1] elaborate the entire process of the quality assessment shown step-by-step from processing data to the newly developed concept of prioritization-based subsamples. The novel methodology developed is shown in Figure 10. This section presents a summary of the methodology and the key concepts and their contributions towards quality assessment methods. The components mainly discussed in this section are highlighted in grey in Figure 10. Particularly tackled here are the geographic information system (GIS) procedures with regards to defining neighbourhood zones and their prioritization to get prioritization-based subsampling strategies (PSSs). Furthermore, the importance of a sample prediction model in the context of developing the quality assessment is explained. The section ends by defining smartly assessed quality of the OSPI service.

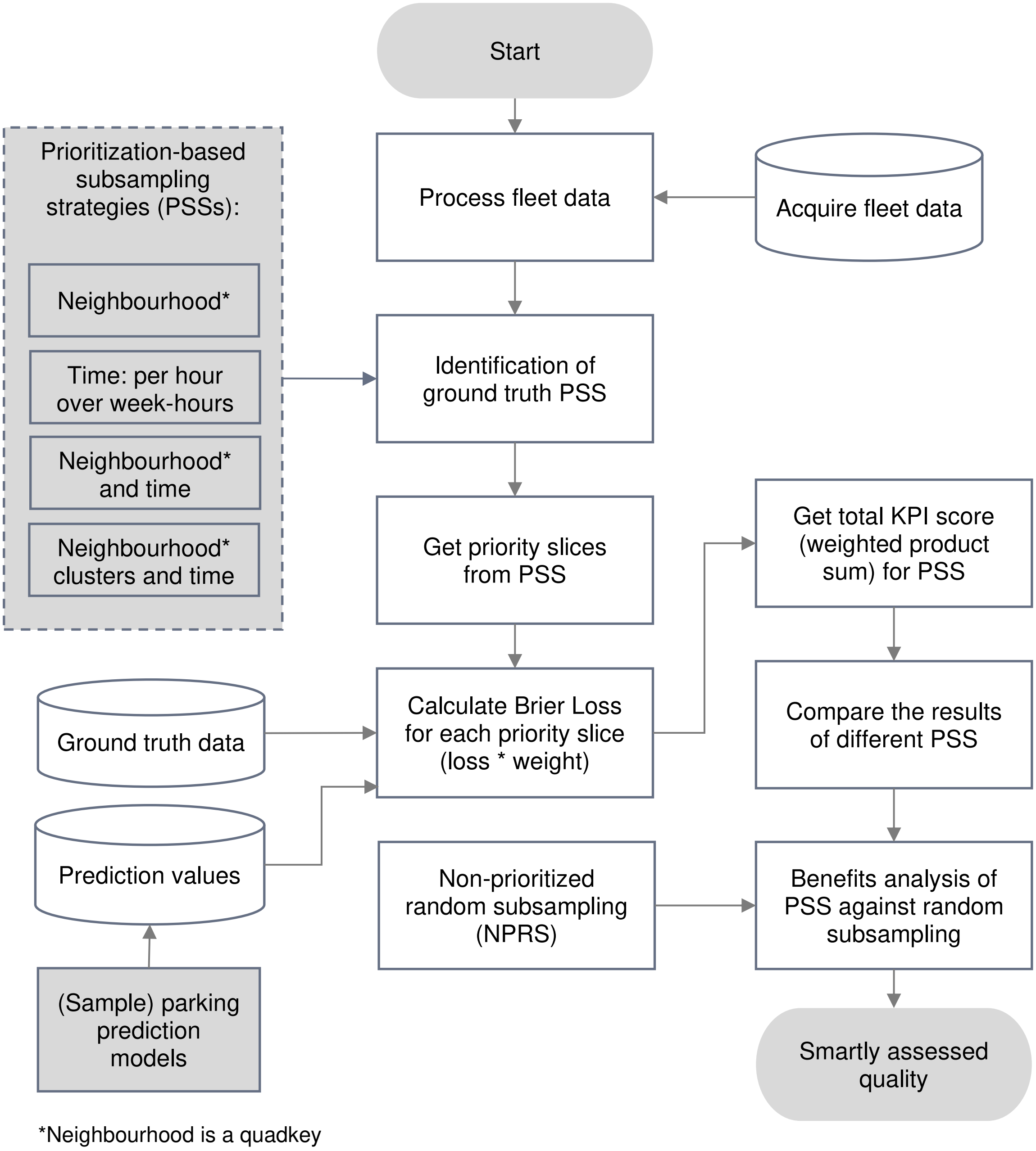


Figure 10. Workflow showing the full development process of the developed quality assessment, based on [1, p. 2]

3.4.1 Development of prioritization-based subsampling strategies (PSSs)

Based on the exploratory analysis done in Gomari et al. [2], it was identified that the TTPD time-series of each quadkey was unique, albeit some having similarities when the volume of parking events data was normalized. However, the analysis clearly showed that the peak hours vary depending on the quadkeys observed. Hence, the approach in the development of the prioritization-based quality assessment [1] hinged on the idea that the importance of the neighbourhoods changes depending on what time of the day is being studied. These areas were more important for ground truth data collection. Gomari et al. [1, p. 2] defined these spatio-temporal segmentation as *slices* and the of each importance is measured based on the "[...] percent volume weight (or density) of fleet data that occur within a certain area and at a specific period, hereafter referred to as *slices*". Having the concept of slices in mind, Gomari et al. [1] proposed a novel way to prioritize neighbourhoods (in terms of quadkeys) and time periods by introducing the so-called prioritization-based subsampling strategies (PSSs) (see the input of this into the entire methodology in Figure 10). These were developed based on the different combinations of slices – in total, there were 10 design setups across all the four strategies. The goal was to understand what level of spatio-temporal segmentation was necessary to achieve a sound quality assessment considering the volume of data available.

In total, four PSSs were developed. The strategies were as follows: (1) based on different zoom levels of neighbourhoods or quadkeys (PSS1); (2) based on time, specifically, 168-week hours (PSS2); (3) based on the combination of neighbourhood and time (PSS3); and based on neighbourhood clusters from the exploratory study [2] and time (PSS4). Among the four, PSS1 stood out as the best strategy for the use case of on-street parking information and using parking events data as the importance data source [1, p. 13]. Specifically, PSS1 with the experimental design based on neighbourhood slices at zoom level 15 (1222.99m x 1222.99m quadkeys). This setup presented the best balance for quantifying importance. The PSSs with time were not great candidates as the 168 week-hour created too many slices even on higher level quadkey zoom levels; this was problematic as there was a lack of ground truth data in comparison to availability of parking events throughout the 168 week-hour. In future work, it is recommended to collect more ground truth data in order to further assess the limitations and potentials of the introduced PSSs for the quality assessment method.

3.4.2 Sample parking prediction model development

Once the PSSs are identified and the slices are calculated (see Figure 10), sample parking prediction models are developed to test and assess the quality against the ground truth data. This section explains further the role of the sample models created for the study. For further details on the specific application example of the prioritization-based quality assessment refer to Gomari et al. [1]. Sample prediction models were built to test the hypotheses of the methodology for the quality assessment. Since developing a full-scale more complex models needed more time (as presented in [4]), sample models were used to verify the benefits of the quality assessment methodology and not necessarily compare the differences in performance of various models. Particularly, the study focused on comparing the different ways of assessing the models instead.

The sample OSPI parking prediction models developed included a combination of random guesser models and actual models using machine learning algorithms (i.e., Xgboost and Random Forest) with real features – although not as extensive as the ones in Chapter 4. In Chapter 4, the development process of a self-adjusting prediction model based on data-driven features is presented. As presented in Gomari et al. [1], the sample models developed only used relatively direct features including temporal features, on-street parking capacity per street, geolocation information with latitude and longitude, engineered features such as basic time-series historic TTPD, and real-time TTPD. As for the random guesser models, four models were created for comparison. First, a completely random model that guesses a probability between 0 (no parking) and 1 (available parking) on a uniform distribution. Second, an optimistic model that guesses between 0.7 and 1.0. Third, a pessimistic random guesser that only guesses between 0 and 0.3. And lastly, an unrealistic model that always makes the best guesses based on the average availability in the entire ground truth dataset. The last model was an interesting case and a model that further validated the prioritization-based methodology. This model assumes the same regardless of the area or time – something completely opposite of a usual driver's behaviour, and even more, a specific group of users like that of BMW car users.

Even before implementation of the methodology, it was expected that all the machine learning-based models and random guessers would perform worse. The reason being that these models do not consider the importance of different locations at different times of the day. It was proven, as shown in [1, Fig. 9], that the methodology does indeed find weaknesses in the models, particularly, in busy areas of the city. The best guesser model was also debunked despite minimizing the brier loss metric in its favour – a weakness this research has reiterated. The only model that showed an improvement in performance was the pessimistic model, since busy areas tend to be occupied in most cases, but not always. In this experiment, it was concluded that, if the methodology is applied in quality assessment, further detailed analysis of the models is needed in identified important areas. Thus, the primary driver to improve a model would be to enhance prediction, specifically, in places that are most frequented by drivers at specific periods of time. This also means deeper investigation is required to understand the features that could model the rare occasions that an on-street parking spot becomes available in busy urban centres.

3.4.3 Smartly assessed quality

The biggest motivation for this dissertation was to find a method to lessen unnecessary or low-impact ground truth data collection. The main goal was to help OSPI achieve a better product-market fit [75] by ensuring the real perceived quality is measured. In the context of quality assessment, this means the OSPI service rendered becomes more useful to the customer. To do that, this research has now developed a methodology that is capable of automatically directing product managers to look at the most important slices (i.e., the areas and time periods) and ensure the utmost customer experience in this period.

As argued in this dissertation, merely using metrics (see Section 2.2) to measure the quality of a prediction model can give a false assessment of a model's performance. And not just that, the common standard in statistics is to do a simple random sampling [89] without doing it strategically. Unfortunately, in the case of mobility-related information services there are spatial and temporal considerations. Moreover, once a service is tailored only for a certain set of users, random sampling may not be the best approach. Gomari et al. [1, p. 11] verified this by comparing the developed prioritization-based methodology against a non-prioritizes randomized subsampling (NPRS) approach. "In the majority of the cases [...] the models performed worse in comparison to NPRS. This implies that assessing the quality at the defined important slices must be checked first before other areas and time periods are observed". This proves that, in majority of important scenarios for customers, randomly selecting ground truth may lead to unfortunate selection of data collection areas and time periods that could lead to a misjudged assessment of the service quality. Especially, when only a metric is considered without any application of correction factors. This potential mistake can be avoided using the introduced prioritization-based methodology, which automatically identifies important slices to initially check and assess the true quality of a mobility-related information system.

3.5 General application in mobility-related information systems

The developed prioritization-based quality assessment is also applicable to other mobility-related information systems. The only requirement is to have big data about the number of users in the system of concern. This can then be used to automatically quantify and identify importance hotspots in the area and keeping in mind that importance changes in space and time depending on demand. Information systems are great agents to not just improve the experience of its users, but also update its services when a prioritization-based quality assessment is employed. In this section, two direct examples are presented: the usage for public transport information systems and car navigation systems.

3.5.1 Public transport traveller information systems

The applications of intelligent transportation systems (ITS) in public transport have significantly improved its services. ITS have allowed automated data collection, which is helpful for planning and operations of public transport systems [90]. Public transport passengers benefit from information at critical times. In this case, prioritization can be on a level important to a traveller. A possibility is to create important slices (see Section 3.4.1 for definition) on a route-level, where the most important routes are ranked. It can also be done on a station-level to determine the most congested stations at peak demand hours. The goal in this use case is to reduce the amount of possible delays in the most critical connections to ensure service reliability and convenience. The method used in Section 3.4.1 employing neighbourhood segmentation using standardised tiles or quadkeys can also be applied for the case of public transportation. It depends on the type of data used for prioritization. For instance, if anonymised smartphone geolocation data is used, the quadkey approach is viable to determine important public transport hubs (i.e., segmented in quadkeys). If traveller information is subpar at these transit areas that could cause delays, but it is perfect in quiet suburban stations, the system will be punished more since less accurate and unreliable information was provided in important slices.

Another issue the methodology can automatically address is during unbalanced ground truth data collection. If there are only 5 ground truth data points during busy hours but 20 during off-peak hours, this is an unbalanced dataset, which will give a high precision score, but not telling the real story behind. If only a metric is used, the score and performance of the models will be assessed as great. When the prioritization-based approach is applied, the methodology automatically adjusts the importance weights to lower the impact of less important regions and amplifies the assessment in areas deemed more important. In the end, the proposed methodology assesses the true quality of a public transport traveller information system.

3.5.2 Car navigation systems

The use case of on-street parking is a primary example as shown in this dissertation. OSPI being part of smarter navigation systems means that the developed methodology is likewise applicable to other services of car navigation. This means the proposed approach is applicable to routing, real-time traffic information, and incident and hazard warning, among others. Since car navigation follows the same structure as the use case applied in this research, the entire method could be replicated with the provision of fleet data relevant for the use case.

Real-time traffic information (RTTI) & Routing. The approach with RTTI is like OSPI. Since GPS tracks from floating car data (FCD) are the primary source of information here, it is clear that prioritization strategies should be geared towards areas with the most congestion. That means forecasting wrongly here it could other dependent services to perform poorly as well. Routing is a direct dependent of the RTTI service. Although it is obvious that congested areas need to be best calibrated for accurate information, quality assessment methodology still do not automatically recognize this issue. In a prioritization-based assessment scenario, overtime, it can be calculated whether the most important areas are receiving better and more reliable information. Indirectly, a key performance indicator here is measuring the number of active users of a system.

Incident and hazard warning. Another service provided in premium navigation systems is incident and hazard warnings. The applicability of this methodology can extend to scenarios involving potholes, damaged road, and road crashes. Such a system provides information about road safety critical areas. Acquiring such low-level information about road conditions, is even more challenging than OSPI. Some systems use crowd-sourced information on such localized information, but more recently autonomous vehicle researches have been using cameras to detect infrastructure anomalies [91]. Another aspect is report road crashes, which are typically reported to the police department and fed into a database that could provide the incident warning systems about road safety concerns in the area. Once a good baseline is identified to implement the prioritization-based methodology, it is expected that with a widespread use of the service, safety indicators will improve overtime. As opposed to a non-prioritization-based strategy, one can use normal floating car data, but this may not match with the users' needs for a hazard warning system.

4 Quality Assurance by Parking Events-Based Data-Driven On-Street Parking Information Systems

Classic parking prediction models mainly use historic parking occupancy data based on manually collected ground truth data or parking sensors in only specific areas of a city. Manual ground truth observations, as repeatedly mentioned, is not a sustainable and scalable method of keeping a prediction model up-to-date. In the case of using on-ground parking sensors, this is usually limited to specific streets and areas, and it is quite costly to construct and maintain due to sustained impact load. Although sensor data contribute to a type of solution that works for some streets, it is not easily replicable in most cities considering budget constraints and different transport policy measures to contain car usage in the city. With the rise of big data in urban mobility, there has been interest to look at data-driven solutions to help keep the quality of an on-street parking information system up-to-date. Gomari et al. [4] propose to use parking events to assure quality in machine learning-based parking prediction models without continuously collecting ground truth data. Figure 11 illustrates their developed methodology to achieve this. For the detailed methodology elaboration refer to [4].

This section primarily focuses on discussing the importance of integrating external data-driven features that are independent of the primary ground truth data. These features differ in their influence on a model. Normal features are dependent on and intrinsic in the ground truth data, whereas data-driven features are detached and act as a second validation dataset relying on continuous data coming in; this includes parking events, where there are many ways to create features that attempts to capture a pattern that predicts the on-street parking state comparable to the traditional ground truth. Additionally, an advantage of the parking events dataset is its capability to detect parking behaviour changes that are not easily identified by costly manual ground truth collectors as presented in [4] – this is summarised in Section 4.1.2. Furthermore, this section provides a brief overview of the key steps in developing a data-driven on-street parking prediction model that aims to self-correct to assure the quality of the OSPI service.

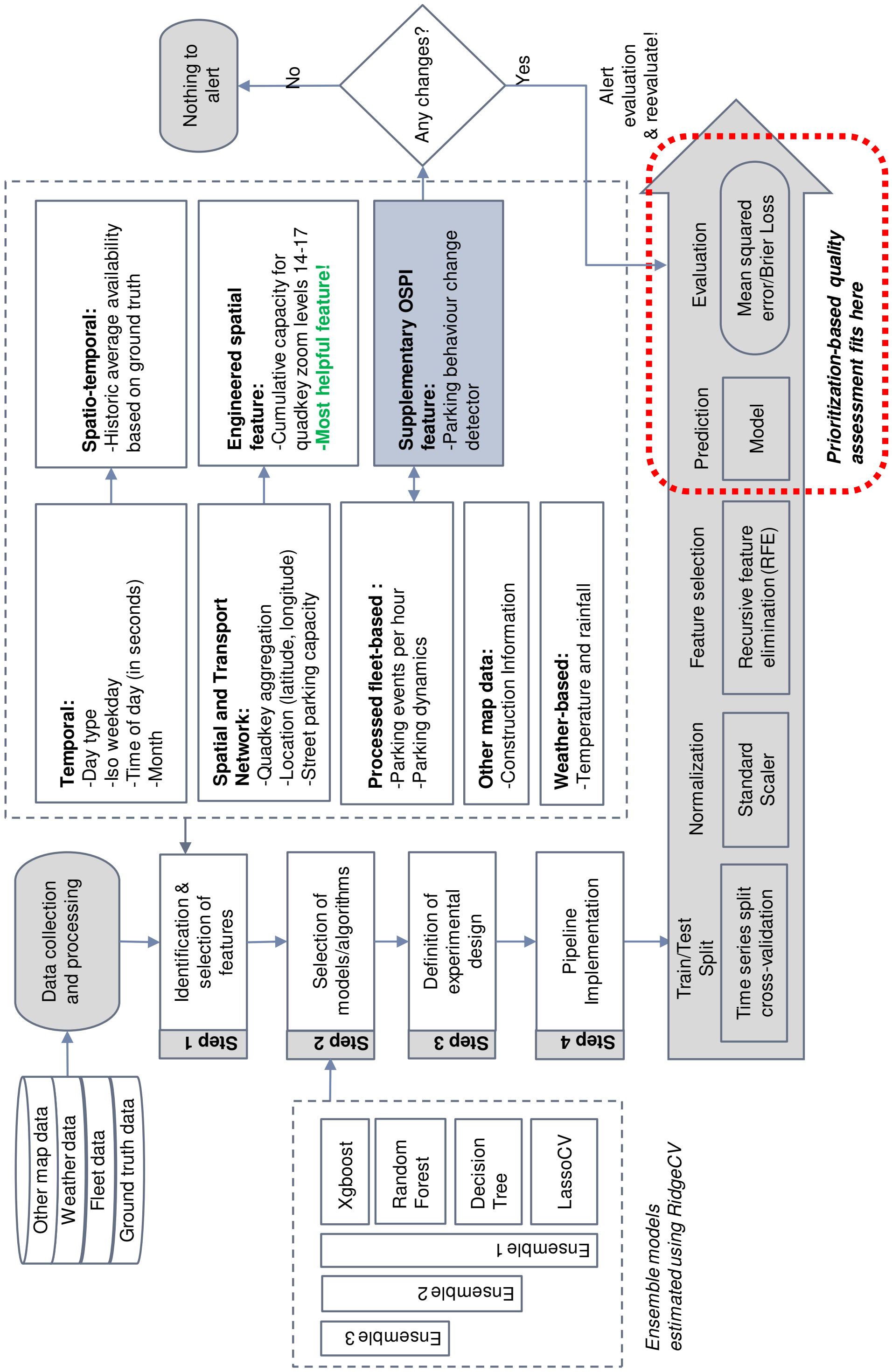


Figure 11. Overview of a data driven OSPI prediction model development, based on [4, Fig. 3]

4.1 Generation of enhanced parking-related features and data-driven supplementary components to assure quality

Gomari et al. [4] have extensively presented the final features that were used in the data-driven parking prediction models developed considering different experimental design setups. In this subsection, the process and significance of the generated features are further discussed in detail and the rationale behind them. The further usage of the developed parking behaviour change model is also summarised at the end of this subsection.

4.1.1 The power of enhanced parking-related features

In comparison with the sample prediction model (see Section 3.4.2), the enhanced parking-related feature models are based on an extensive set of features engineered and filtered through a feature selection step – that is using Recursive Feature Elimination (RFE) as implemented in scikit-learn [92]. As opposed to limited engineered features as it has been done in existing literature, the aim of this study was to further generate features from existing data sources that could possibly capture more variances to better predict OSPI availability. This was possible because of the availability of parking events data as secondary validation that feeds into the model as a set of features. As shown in Table 3, Gomari et al. [4] generated 102 features in total to find features capable of generalising towards the target value prediction. Based on empirical evidence from the trials, in the end, only 21 features were selected in the final models. Below are why and how the features were generated (see Table 3).

Spatial & transport network features. Seven features were generated to cover this. Features considered here are those inferring from geolocation position of an area or event, and usage of spatial parking capacity. This set of features turned out to be the most vital. In majority of experimental design setups, spatial-related features always captured the highest variation as part of the estimators – this may also be the reason for the neglecting their enhancement. Thus, Gomari et al. [4] tested this with different segmentation strategies using various quadkey zoom levels to aggregate on-street parking capacity. This means on zoom level 14 the total capacity over a 2.5km x 2.5km tile is considered. The test showed, simply by transforming street-level capacity to neighbourhood-level capacity, the prediction model was able to enhance generalisation of situation, while also considering the GPS geolocation position. Refer to [4, Fig. 10] to see the importance of each feature, where experimental design (ED) setups with spatial features demonstrate the significant influence of the introduced features.

Temporal features. Time-related features were the second most important set of features with 15 features tested. The process here was to generate various aggregation of time. The different levels included: month, week, day, hour, hour of the day, at what minute of the day (in minutes), at which 15 or 30-minute interval during the day (in 15 or 30-minute intervals), second of the day (in seconds), weekdays, weekends, holidays, at which hour in the week (in 168 week-hours), and days from or after holidays, among others. In comparison to the benefits gained from the spatial features, the impact of the temporal features was slightly lower. The top 2 features as shown in the model of interest ED5 in [4, Fig. 10] are time of day in seconds and week-hour. This demonstrates that these two features are the best estimators when considering predictions over a long-term horizon.

Weather-related features. This category was quite simple as it only considered temperature and rainfall. Although these two features could be beneficial for special extreme events that lead to disruption, they were insignificant in generalisation as estimators for an OSPI parking prediction model.

Historic parking availability. Among all features, historic availability is the most used in almost all studies. Nine features were created related to historic parking availability. This simply depends on the ground truth data and checking for ways to generalise over different moving average periods. Essentially, the goal in historic data related features is to understand patterns and the recurrence; this is the reason many studies attempt to use LSTM recurrent neural networks to capture the patterns. Since parking prediction studies are a supervised learning problem, and there is a wide range of empirical findings in relation to feature generation, a neural network approach may not be the best option; as also proved in few studies, especially, considering the processing time performance. As shown in Figure 11, aggregation of historic parking availability was done on different quadkey levels and different calendar week moving averages. Overall, the set of historic moving averages features was proved to improve the models [4, p. 6] when all feature categories were involved. However, when historic averages were added to the models with temporal, spatial, and weather features, the performance slightly dipped. The enhancement was only observed when it was combined with parking events-based features as discussed next.

Parking-events based features. The aim of generating a variety of enhanced features was to reduce the reliance on historic availability information after the model training period at the launch of an OSPI service. Gomari et al. [4, p. 6] proved that features engineered from parking events can replace historic ground truth information. As shown in the scores table in [4, p. 6], on average, ED5 models perform well even without historic averages. This shows that parking events features can indeed be a replacement for a given period (i.e., within a 3-month horizon as done in the train and test split [4]) without requiring updates from ground truth data. Before this was achieved, 69 parking events-based features were tested. They are broken down into three categories: (1) based on historic parked-in or parked-out events, (2) historic and real-time temporal trend of parking dynamics (TTPD), and (3) parking behaviour change detection (PBCD). The first two categories are enhanced by combinations of aggregations on various moving average intervals and different quadkey or tile zoom levels, as mentioned for the spatial features; while the last category is discussed in Section 4.1.2. In its main findings, Gomari et al. [4, p. 9] mentioned: "5 out of top 10 most important features for the mentioned model are parking events-related". This is proof that parking events-based features can supplement parking prediction models to reduce reliability on entirely manual ground truth collection. Parking behaviour change detection did not have any impacts on parking prediction models. PBCD benefits are primarily noticeable for events that affect predictions in the longer term and not real-time. And since not many grand changes were seen in the city of Munich, this could not be tested; also, considering that usually, ground truth is not being collected when disruptions happen.

Table 3. Features engineered for enhance OSPI models, based on [4, p. 5]

Feature category	Number of Features	Description of feature content
Spatial & Transport Network	7	GPS location, on-street parking capacity features divided or aggregated on different spatial levels
Temporal	15	Only time-related features considering aggregations into time intervals in different time scales and categorization of special days: months, weeks, days, hours, minutes, seconds, weekdays, weekends, holidays, etc.
Weather	2	Rain and temperature open data
Parking availability	9	Aggregation of historic parking availability on different tile levels and time intervals (e.g., moving averages) in the past.
Parking events-based	69	Automated aggregation in various time intervals of temporal trend of parking dynamics (TTPD) [2] that describe on-street parking activity on tile zoom level 14, and aggregation in various time intervals (e.g., real-time and moving averages) and tile levels of parked-in and parked-out events; anomalies detected based on the developed behavior change detection in

The other model implementation processes and feature selection are described in Section 4.3.

4.1.2 Parking behaviour change detection: Data-driven supplemental OSPI component

State-of-the-art OSPI systems as shown in literature have focused on maximizing the usage of ground truth to train the models. The idea of utilizing other sources to trigger updates or directed strategic ground truth collection has not been explored. This is not surprising, as on-street parking-specific geolocation data is not easily available for research. The parking events dataset used in this dissertation is a unique data source that has many use case potentials for improving OSPI services.

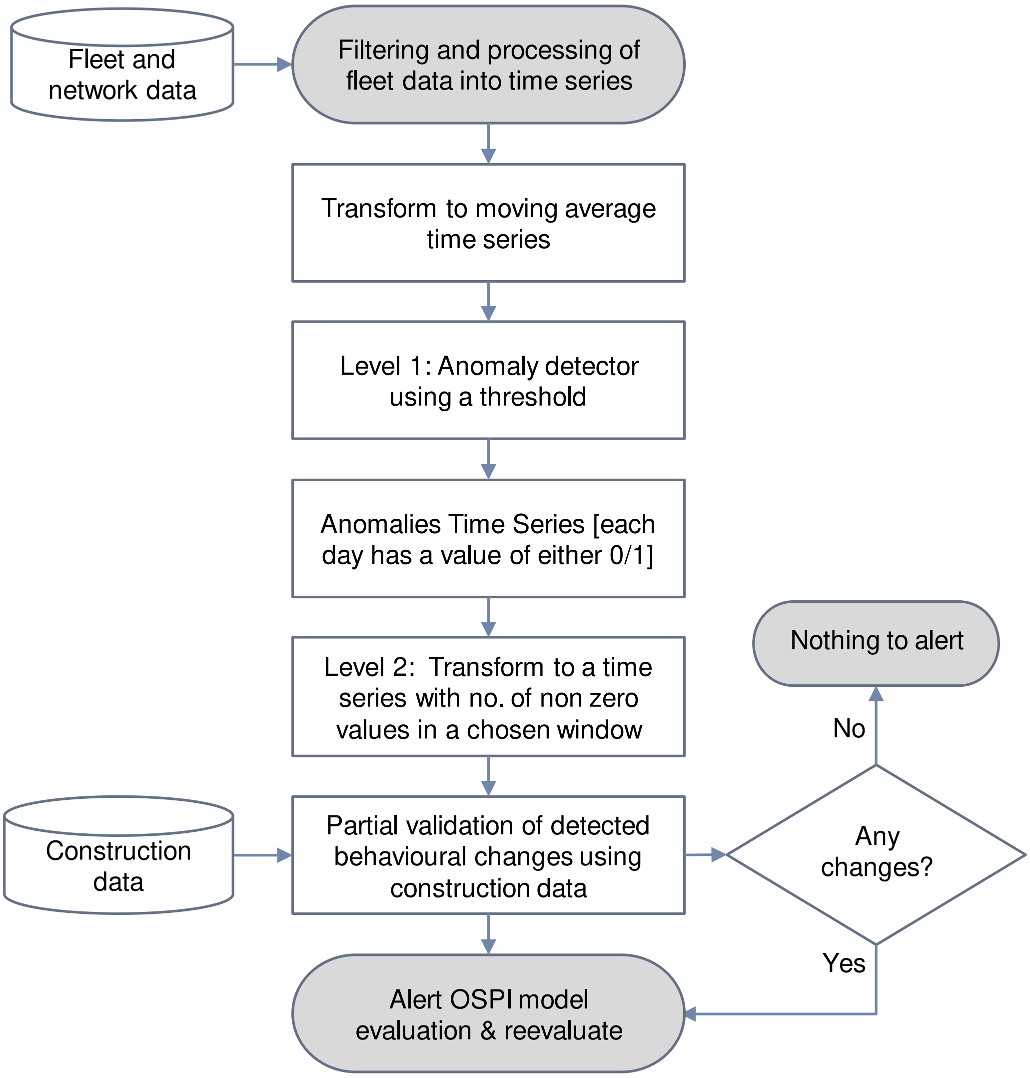


Figure 12. The parking behaviour change detection model, as presented in [4, p. 11]

Gomari et al. [4] (see Figure 12) and Domakuntla [51] have proposed methodologies to utilise parking events as update trigger systems, specifically, for parking behaviour change detection (PBCD). The methodology consists of a two-level anomaly detection model primarily created to detect disruptions that lead to on-street parking blockage such as in the scenario of construction. As illustrated in Figure 12, the parking events data is transformed into a smooth time series format with moving averages to fill potential gaps. Then, a lower threshold is set relative to the maximum value in the time series in order to detect complete disruptions. This is then transformed to another time series of whether the value was above or below the threshold. The next filter checks for how long the disruption lasts. Once it also satisfies this criterion, then it is considered an anomaly given the indicated time window. Gomari et al. [4, p. 12] further validated this using construction data from HERE maps and where available also checked with open data from the city of Munich's website. It was determined that implementing such a feature is currently viable to keep the OSPI service up-to-date.

It must be noted that, the volume of parking events generated has yet to reach a significant penetration level to influence real-time parking prediction systems. However, as shown in Gomari et al. [4] study, the volume is currently sufficient to detect medium (a few weeks) to long term (months) changes. This is a significant finding considering that, the usual map updates for real-time traffic information or routing (i.e., more mature services) get updated only 4 times a year [73]. Integrating such a trigger function into an OSPI product will serve as a supplementary component that can assure quality besides collecting manual ground truth collection.

4.2 Hyperparameter tuning process of the OSPI prediction models

This section discusses the setup and the taken steps to tune the hyperparameters of the implemented models in the study of Gomari et al. [4]. For a detailed analysis of the parking prediction results refer to [4, p. 5]. The machine learning algorithms or models selected were based on the literature review conducted as presented in Section 2.2. The commonly performant models were selected for comparison and the ones proven inferior in most studies were left out to reduce comparison parameters. Besides, a goal in this dissertation was to utilise algorithms that are rather simple to implement in comparison to hybrid complex models, which can dilute the analysis and interpretation of the models developed.

4.2.1 Models selected

Based on the literature conducted by Gomari et al. [4, p. 4], the most performant machine learning models in OSPI are primarily tree-based models: Xgboost, Random Forests, and Decision Tree. The most basic of them three is Decision Tree. Decision

Trees split the observations based on the most common feature at each node created until the last possible split, where the subgroups that have been created are as similar as possible. Extreme Gradient Boosting or Xgboost [61] is essentially an ensemble model that, sequentially, builds trees and predicts. Xgboost learns from the mistakes of weaker decision trees and corrects these predictions. Xgboost's greatest attribute is its fast yet accurate prediction capability because of its simpler structure relative to neural networks, for instance. Random Forests [63], [93] are simply composed of a large number of decision trees that are built in parallel that are based on bagging (i.e. random sample selection) and feature randomness (i.e. selecting random set of features). The principle followed is that with many uncorrelated trees, that have their own random set of features and training sample, mistakes would be made, individually, while the outcome is decided based on the average of probabilities from the predictions made at the end. Furthermore, apart from the chosen models, deep learning models in combination with gradient boosting tree models have shown better performance in some instances, however, "given similar performance in comparison to efforts in long processing time" [4, p. 4] and model complexity, these models were left out in experiments.

4.2.2 Tune hyperparameters

A crucial step in model implementation is conducting the hyperparameter tuning. The essence of this is to automate parameter changes that are input in the models implemented. Two methods were used in the study of [4, p. 5]: (1) exhaustive grid search, also called GridSearchCV in scikit-learn [92] and (2) randomised parameter optimisation, also called RandomizedSearchCV as implemented in scikit-learn [92].

Among the three models used, the fastest algorithm to tune was the Decision Tree models, shortly followed by Xgboost, and Random Forest took the longest. However, the most difficult to tune was Xgboost since it had the greatest number of parameters to tune. Since runtime with Decision Tree and Xgboost were relatively fast, grid search was used in both cases. The initial step here was based on literature and heuristics to define initial range of values to test. As a start, this can be as low as two initial values that are far apart for each parameters and observe the difference in performance of the models. The next step is to adjust the range of parameters. After a few iterations, it is all about optimisation – and here, a low interval step is recommended, until there is no significant changes in the mode score. It is also best practice to tune the parameters one by one for more precision and avoid going back and forth between parameter tuning.

It must be noted that, each model could have special parameters that have to be adjusted depending on the problem in hand. For instance, in the Xgboost implementation, it was important to change the objective parameter to "binary:logistic" since the target value was probability of parking availability, which is between 0 and 1. For ensemble models such as Random Forests, the parameters were optimised for the best results, but since it is also an ensemble model that combines many individual trees, the parameter tuning did not have as much of an impact compared to the impacts from features engineered. Hyperparameter tuning was discovered to be, however, more important in a basic implementation of Decision Tree models. This is because the default parameters from the library are the bare minimum to allow implementation. Gomari et al. [4, p. 6] for instance observed a 31% improvement in the model implementation since the min_samples_leaf and maximum depth were adjusted. Nonetheless, Decision Tree models were not performant in any of the scenarios, but it was a good baseline since the other two ensemble models are based on it.

4.3 Other key learnings from developed OSPI parking prediction models

4.3.1 Experimental design setup

A major step before model implementation is designing the experimental design parameters and the combinations thereof to test different model scenarios. In the study of Gomari et al. [4], an incremental heuristic approach was taken to determine the set of combinations based on the feature categories (see Table 3). Experimental design setups are necessary to understand which combination of features and/or model algorithm render the best performance. In the study of Gomari et al. [4], the most important learning was that parking events-based features could replace industry historic parking availability features and marginally outperform them. This is a significant finding as incorporating parking events-based features can lessen the dependency on continuous ground truth collection. Additionally, in combination with the parking behaviour change detection (PBCD) model, the OSPI system is alerted with triggers for possible mid to long term disruptions in parking availability in an area.

4.3.2 Feature selection with Recursive Feature Elimination (RFE)

As mentioned in Section 4.1, many features were engineered to explore which set could capture the most variance to improve predictions. Since many features were generated, instead of doing the selection manually, Recursive Feature Elimination (RFE) was used as implemented in scikit-learn [92]. There are also a variety of methods for feature selection, which can change depending on the use case. The RFE approach recursively reduces the set of features used in the model, while pruning the least important ones in each round.

Once implemented in the study, from 102 features generated, only 21 were retained in each model implementation, as the rest could not even capture more than 1%

importance factor (i.e., based on variance), if any at all. An important finding was that a lot of features are not needed for regression problems since the requirement is to have less features that can generalise better. If there are a lot of features included but only contribute marginally, this could lead to overfitting, which is not desirable in a use case such as OSPI, where scenarios and contexts can differ from one another.

In this study, the ideation stage was the most difficult part. It is also important to consider that feature generation has a bias towards the background knowledge of the data scientist dealing with the problem. It was discovered that, in general, a good balance between aggregated features and detailed features is needed. A feature aggregated on a high level may not be able to distinguish changes, while a local or detailed feature may result to overfitting and not able to predict properly – a downfall of many machine learning implementations, if not done cautiously.

5 Discussion and Future Research

Following the summaries of the three main studies discussed in Chapters 3 and 4, this section follows up with a discussion of constraints and the direction of future research. A list of suggested future studies is also elaborated here.

5.1 Data collection constraints

A great advantage of this dissertation was the availability of some data sources within the first year of research. A qualitative assessment was done (see Section 2.1) comparing the different available datasets. However, early analysis showed that there were certain aspects beyond the control of the researcher. For instance, changing the data collection specifications by the fleet – as it is a business process within BMW, and cannot be easily changed, and was also not in the scope. Nonetheless, the goal was to explore the possibilities with the existing data that was being collected at the time. Furthermore, a promising solution at the beginning of this research was data fusion; however, because of the disaggregated nature in collection of the different sources, and the other complexities involved in post-processing, it was decided to stick to one reliable source rather than many sources that are difficult to combine and unavailable. Another limitation was collecting manual ground truth data, as it was a costly process. A few studies for validation of experiments were supported, and regular random collection was done, but doing strategic data collection could not be done on a regular basis due to financial constraints.

Like most studies, the data availability for this research was also impacted by the COVID-19 pandemic. Since behaviours changed and some confidential data specifications and projects in the horizon changed, the data used in this dissertation were bound by a period of study. Nonetheless, this was not a hindrance to the research, but rather a missed opportunity due to an unforeseen pandemic.

5.2 Limitations and further extension of the assessment methodology

The prioritization-based quality assessment methodology developed (see Chapter 3) in this research does not include software and system quality assessment. This refers to the impacts related to user interface and user experience aspects. The developed automated system also did not focus on the potential impact of missing data, where an alert trigger system could be built to alert when there are no ground truth data at important areas. This is a minor detail that could be extended in future extensions.

As there was only limited time to develop the methodology within this dissertation, there is a lack of post monitoring after application on the proposed OSPI prediction models. To extend this part of the research, after conducting the prioritization-based assessment, further ground truth data needs to be collected to re-evaluate and analyse changes in model scores.

To further extend the quality assessment, it is suggested to apply the methodology on a higher level that assesses other mobility-related information systems as well, such as real-time traffic information. This will make the methodology easier to implement since adjustment will be made to full tailor the methodology to all the products that are crucial for smarter navigation systems.

5.3 Recommended research direction of future on-street parking information systems studies

OSPI systems will continue to play a role in urban mobility. The shift towards more automated and connected systems will shift the information systems as necessary and proactively. The ability of cars to communicate and cooperate their search for parking spaces can lessen congestion and improve information systems on a city level. Nonetheless, prediction models to estimate open spots will still serve in the background as a guide from the origin. It is foreseen that ground truth collection will move towards assisted systems with computer vision.

In general, future research should focus on data fusion to enhance mobility-related information systems and particularly, OSPI. Being a multivariate problem, on-street parking information accuracy can benefit enormously from data fusion by using the advantages of different data sources. Fusing of data will also enhance the proposed prioritization-based quality assessment, as a higher volume of data is available for prioritization. Camera as sensors will be the future direction of research combined with redundancies from data sources that will validate them. These data sources can include parking events, municipality parking information from street sensors, ultrasonic sensors, and LIDAR data that is mostly used in autonomous driving.

Three main topics were tackled in this dissertation including cluster analysis of parking behaviour, prioritization-based quality assessment, specifically, for OSPI, and development of data-driven parking prediction model. In relation to the mentioned future overview and studies already conducted, here below is a list of topics and research questions that can be tackled within the scope of OSPI and mobility-related information systems, in general:

- As suggested by Gomari et al. [1, p. 11], a next possible research direction for the prioritization-based quality assessment is comprehensively studying and identifying the optimal minimum fraction of ground truth required for the proposed true quality assessment check. To do this, an extensive amount of random and strategic ground truth data collection will be needed.
- The application of the prioritization-based quality assessment methodology on other mobility-related information use cases, and "the extension of prioritization-based subsampling strategies (PSSs) using other factors such as the density of points-of-interest (POIs) or local contextualized information and so on" [1, p. 11].
- Current OSPI prediction models still focus on localised availability prediction per street. However, this does not completely represent a driver's behaviour in search for on-street parking. Drivers typically go to an area and knowing their chances (in temporal terms) to find a parking spot. It is suggested to conduct a deeper investigation to understand the features that could potentially model the rare occasions that an on-street parking spot becomes available in busy urban centres. For this, focused data collection and research on limited parts of a city is the first step. Direct interviews and understanding drivers' needs and requirements are also recommended to understand the gap in the market. This study can be combined with parking search route research work.
- Parking behaviour studies not related to prediction models are still uncommon. As argued in Gomari [2], it is crucial to understand parking behaviour to create meaningful features for prediction models. Gomari et al. [2, p. 8] in relation to their study, suggest to extensively research on activities specifically related to on-street parking choices. This can be done in the form of stated and revealed preference studies. And as a supplementary validation dataset, cleaned OpenStreetMap[4] points of interest (POIs) can be used to correlate with potential activities of users. As mentioned in Gomari et al. [2, p. 8]: "This helps better comprehend possible correlations between quadkeys and an activity performed resulting to more localised estimations and the detection of popular areas in time and space".
- As technology has evolved over the last three years, a follow-up research to address the data collection constraint is to use computer vision with cameras to collect on-street parking ground truth. Particularly, develop a spatio-temporal

[4] https://www.openstreetmap.org/

collection strategy that could complement the prioritization-based quality assessment of Gomari et al. [1]. The focus of this research should be on comparing the benefits gained shifting from manual collection to automated detection from organic crowd-sourced camera or fleet camera data. It is also useful to estimate the error rate and the correction needed to compensate for the errors.

- More related to the product side research, a recommended study would be getting popup user feedback in the navigation system for quality feedback. The feature can be embedded into the system. This is direct valuable feedback that could be beneficial both for the customers and the service provider, in this case the car OEM. This study will be supplementary research to the existing system in place.
- In all the studies conducted in this dissertation, the quadkey approach was used to make reproducibility viable. Another approach that can be taken in future research is the application of gradual rasterization as defined by [94]. The proposed method would be to conduct an analysis and adjust study area size depending on the amount of volume on the tiles, quadkeys, or zones. This means, geographically small but high-volume tiles can be compared with large tiles with the same amount of volume. This has already been done for transport modelling studies like those in [94], [95]. The focus of this study should be comparing the results and identify potential gaps in opportunities to do better prioritization-based quality assessment.

6 Conclusions

On-street parking information (OSPI) systems will continue to play a major role in the context of smart cities. This dissertation has extensively studied and developed methodologies to assess the true quality of a mobility-related information system with a focus on OSPI as presented in the studies of Gomari et al. [1], [2], [4] as presented in Appendices A, B, and A. Quality assurance, typically, is aligned with user satisfaction and product usage and growth. This means, a wider and better usage of mobility-related information systems like OSPI can achieve transport network system-level benefits [1]. Combined with sound sustainable urban mobility policies, OSPI could improve the overall mobility situation in cities today and in the future with more connected, automated, and cooperative systems in sight.

In summary and conclusion, the developed novel prioritization-based assessment methodology for mobility-related information systems, specifically OSPI, has been proven to deliver valuable results as presented in [1], [2], [4]. Even with low volume parking event counts in different cities, the volume can still steer the quality assessment towards areas that are most frequented by customers, in this case, that of BMW. Since fleet data is typically a representation of a bigger population, as demonstrated in [2], it is worthwhile to investigate the users' parking behaviour which can be obtained from the same fleet data. Furthermore, the parking events data, as the core element of the entire dissertation, has proven that reduction of dependence on manually collected ground truth data is possible [4]. Gomari et al. [4] developed a series of models aiming to find the best combination to reduce dependency on manual ground truth data by introducing data-driven features based on parking events and enhanced spatial features. The experiments in the study prove that, even up to a horizon of 3 months as tested in the cross-validation test set, the prediction scores were still performant. And on top of that, a parking behaviour change detection model is proposed that can be used a trigger component to alert the system about potential long-term changes in the OSPI service. Of course, the proposal cannot beat real-time information, but the system is a big step towards reducing reliance on historic ground truth that is costly to collect.

Bibliography

[1] S. Gomari, C. Knoth, and C. Antoniou, “Prioritization-based subsampling quality assessment methodology for mobility-related information systems,” *IET Intelligent Transport Systems*, vol. 16, no. 5, pp. 602–615, 2022, doi: 10.1049/itr2.12160.

[2] S. Gomari, C. Knoth, and C. Antoniou, “Cluster analysis of parking behaviour: A case study in Munich,” *Transportation Research Procedia*, vol. 52, no. 2021, pp. 485–492, 2021, doi: 10.1016/j.trpro.2021.01.057.

[3] S. Gomari, C. Knoth, and C. Antoniou, “Spatio-temporal analysis of parked-in and parked-out events : A case study of Munich,” 2020.

[4] S. Gomari, R. Domakuntla, C. Knoth, and C. Antoniou, “Development of a data-driven on-street parking information system using enhanced parking features,” *IEEE Open J. Intell. Transp. Syst.*, vol. 4, pp. 30-47, 2023, doi: 10.1109/OJITS.2023.3235898

[5] K. Gkolias and E. I. Vlahogianni, “Convolutional Neural Networks for On-Street Parking Space Detection in Urban Networks,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 12, pp. 4318–4327, 2019, doi: 10.1109/TITS.2018.2882439.

[6] H. S. Jomaa, J. Grabocka, and L. Schmidt-thieme, “A Hybrid Convolutional Approach for Parking Availability Prediction,” in *2019 Internationl Joint Conference on Neural Networks (IJCNN)*, 2019, pp. 1–8. doi: 10.1109/IJCNN.2019.8852400.

[7] A. J. Pel and E. Chaniotakis, “Stochastic user equilibrium traffic assignment with equilibrated parking search routes,” *Transportation Research Part B: Methodological*, vol. 101, pp. 123–139, 2017, doi: 10.1016/j.trb.2017.03.015.

[8] Y. Rong, Z. Xu, R. Yan, and X. Ma, “Du-Parking : Spatio-Temporal Big Data Tells You Realtime Parking Availability,” in *KDD ’18: The 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, August 19-23, 2019, London, United Kingdom.*, 2018, pp. 646–654.

[9] T. Friedrich *et al.*, “Routing for on-street parking search using probabilistic data,” *AI Communications*, vol. 32, no. 2, pp. 113–124, 2019, doi: 10.3233/AIC-180574.

[10] R. C. Hampshire and D. Shoup, “What Share of Traffic is Cruising for Parking?,” *Journal of Transport Economics and Policy*, vol. 52, no. Part 3, pp. 184–201, 2018.

[11] K. Kim and N. Koshizuka, "Data-driven Parking Decisions: Proposal of Parking Availability Prediction Model," in *2019 IEEE 16th International Conference on Smart Cities: Improving Quality of Life Using ICT & IoT and AI (HONET-ICT)*, 2019, pp. 161–165. doi: 10.1109/HONET.2019.8908028.

[12] B. Gyergyay, S. Gomari, M. Friedrich, J. Sonnleitner, J. Olstam, and F. Johansson, "Automation-ready framework for urban transport and road infrastructure planning," *Transportation Research Procedia*, vol. 41, pp. 88–97, 2019, doi: 10.1016/j.trpro.2019.09.018.

[13] M. Braibanti, "Our Newest Innovation: Ultrasonic Sensor Parking Availability Technology," *INRIX*, 2017. https://inrix.com/blog/ultrasonic-sensor-parking-availability-technology/ (accessed Feb. 02, 2022).

[14] A.-M. Zambanini, Sebastian, Loghin, N. Pfeifer, E. M. Soley, and R. Sablatnig, "Detection of Parking Cars in Stereo Satellite Images," *Remote Sensing*, vol. 12, no. 13, pp. 1–24, 2020, doi: https://doi.org/10.3390/rs12132170.

[15] Y. Purahoo, G. Gopal, and N. Chooramun, "A Smart Parking and Mapping System Using Aerial Imagery," in *Information Systems Design and Intelligent Applications*, 2019, pp. 421–430.

[16] M. Dzivjak and D. Fiedler, "Estimating the Parking Capacity in Cities Using Aerial Images," Czech Technical University in Prague, 2019.

[17] V. Coric and M. Gruteser, "Crowdsensing maps of on-street parking spaces," *Proceedings - IEEE International Conference on Distributed Computing in Sensor Systems, DCoSS 2013*, pp. 115–122, 2013, doi: 10.1109/DCOSS.2013.15.

[18] S. Gomari, C. Knoth, and C. Antoniou, "Comparative Data Analysis for Quality Assessment of On-Street Parking Information (OSPI)," 2019.

[19] L. Mannini, E. Cipriani, U. Crisalli, and A. Gemma, "FCD data for on-street parking search time estimation," *Iet Intelligent Transport Systems*, vol. 12, no. 7, pp. 664–672, 2018, doi: 10.1049/iet-its.2017.0223.

[20] O. Cats, C. Zhang, and A. Nissan, "Survey methodology for measuring parking occupancy: Impacts of an on-street parking pricing scheme in an urban center," *Transport Policy*, vol. 47, no. April 2016, pp. 55–63, 2016, doi: 10.1016/j.tranpol.2015.12.008.

[21] A. Sunderrajan, V. Viswanathan, W. Cai, and A. Knoll, "Traffic state estimation using floating car data," *Procedia Computer Science*, vol. 80, pp. 2008–2018, 2016, doi: 10.1016/j.procs.2016.05.521.

[22] F. Bock and S. Di Martino, "How many probe vehicles do we need to collect on-street parking information?," *5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems, MT-ITS 2017 - Proceedings*, pp. 538–543, 2017, doi: 10.1109/MTITS.2017.8005731.

[23] 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, doi: 10.1109/TITS.2015.2428705.

[24] S. Yang, W. Ma, X. Pi, and S. Qian, "A deep learning approach to real-time parking occupancy prediction in spatio-temporal networks incorporating multiple spatio-temporal data sources," *Transportation Research Part C*, vol. 107, pp. 248–265, Jan. 2019.

[25] F. V Monteiro and P. Ioannou, "On-Street Parking Prediction Using Real-Time Data," in *2018 IEEE 21th International Conference on Intelligent Transportation Systems (ITSC)*, 2018, pp. 2478–2483.

[26] W. Shao, Y. Zhang, B. Guo, K. Qin, J. Chan, and D. Flora, "Parking Availability Prediction with Long Short Term Memory Model," *International Conference on Green, Pervasive, and Cloud Computing (GPC)*, pp. 124–137, 2018.

[27] F. Shi, D. Wu, Q. Liu, Q. Han, and J. A. Mccann, "ParkCrowd: Reliable Crowdsensing for Aggregation and Dissemination of Parking Space Information," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 11, pp. 4032–404, 2018, doi: 10.1109/TITS.2018.2879036.

[28] A. Origlia, S. Di Martino, and Y. Attanasio, "On-Line Filtering of On-Street Parking Data to Improve Availability Predictions," in *2019 - 6th International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)*, 2019, pp. 1–7. doi: 10.1109/MTITS.2019.8883375.

[29] A. Ionita, A. Pomp, M. Cochez, T. Meisen, and S. Decker, "Where to Park? Predicting Free Parking Spots in Unmonitored City Areas," in *WIMS '18: 8th International Conference on Web Intelligence, Mining and Semantics*, 2018, vol. 88. doi: 10.1145/3227609.3227648.

[30] G. Jelen, V. Podobnik, and J. Babic, "Contextual prediction of parking spot availability: A step towards sustainable parking," *Journal of Cleaner Production*, vol. 312, Aug. 2021, doi: 10.1016/j.jclepro.2021.127684.

[31] D. Zhao *et al.*, "MePark: Using Meters as Sensors for Citywide On-Street Parking Availability Prediction," *IEEE Transactions on Intelligent Transportation Systems*, 2021, doi: 10.1109/TITS.2021.3067675.

[32] S. Di Martino and A. Origlia, "Exploiting recurring patterns to improve scalability of parking availability prediction systems," *Electronics (Switzerland)*, vol. 9, no. 5, pp. 1–19, 2020, doi: 10.3390/electronics9050838.

[33] E. H.-C. Lu and C.-H. Liao, "Prediction-based parking allocation framework in urban environments," *International Journal Of Geographical Information Science*, vol. 34, no. 9, pp. 1873–1901, Sep. 2020, doi: 10.1080/13658816.2020.1721503.

[34] X. Xiao, Z. Jin, Y. Hui, Y. Xu, and W. Shao, "Hybrid spatial–temporal graph convolutional networks for on-street parking availability prediction," *Remote Sensing*, vol. 13, no. 16, pp. 1–20, 2021, doi: 10.3390/rs13163338.

[35] H. Canli and S. Toklu, "Design and Implementation of a Prediction Approach Using Big Data and Deep Learning Techniques for Parking Occupancy," *Arabian Journal for Science and Engineering*, 2021, doi: 10.1007/s13369-021-06125-1.
[36] J. C. Provoost, A. Kamilaris, L. J. J. Wismans, SanderJ. der Drift, and M. Van Keulen, "Predicting parking occupancy via machine learning in the web of things," *Internet of Things*, vol. 12, Dec. 2020, doi: 10.1016/j.iot.2020.100301.
[37] W. Zhang, H. Liu, Y. Liu, J. Zhou, and H. Xiong, "Semi-supervised hierarchical recurrent graph neural network for city-wide parking availability prediction," in *AAAI 2020 - 34th AAAI Conference on Artificial Intelligence*, 2020, pp. 1186–1193. doi: 10.1609/aaai.v34i01.5471.
[38] K. Ivan, T. Marta, M. Kostadin, and T. Dimitar, "Parking Availability Prediction Using Traffic Data Services," in *Conference: ICT Innovations 2020*, 2020, vol., no. September, pp. 1–15.
[39] L. Xiangdong, C. Yuefeng, C. Gang, and X. Zengwei, "Prediction of short-term available parking space using LSTM model," in *14th International Conference on Computer Science and Education (ICCSE 2019)*, 2019, pp. 631–635.
[40] T. Ling, X. Zhu, X. Zhou, and S. Wang, "ParkLSTM: Periodic Parking Behavior Prediction Based on LSTM with Multi-source Data for Contract Parking Spaces," in *International Conference on Wireless Algorithms, Systems, and Applications*, 2021, vol. 12938, pp. 262–274. doi: 10.1007/978-3-030-86130-8_21.
[41] B. Xu, O. Wolfson, J. Yang, L. Stenneth, P. S. Yu, and P. C. Nelson, "Real-time street parking availability estimation," *Proceedings - IEEE International Conference on Mobile Data Management*, vol. 1, pp. 16–25, 2013, doi: 10.1109/MDM.2013.12.
[42] M. Arab and T. Nadeem, "MagnoPark - Locating On-Street Parking Spaces Using Magnetometer-Based Pedestrians' Smartphones," in *2017 14th Annual IEEE International Conference on Sensing, Communication, and Networking, SECON 2017*, 2017, no. July 2020. doi: 10.1109/SAHCN.2017.7964915.
[43] E. G. Mantouka, P. Fafoutellis, and E. I. Vlahogianni, "Deep survival analysis of searching for on-street parking in urban areas," *Transportation Research Part C: Emerging Technologies*, vol. 128, p. 103173, Jul. 2021, doi: 10.1016/J.TRC.2021.103173.
[44] L. Stenneth, O. Wolfson, B. Xu, and P. S. Yu, "PhonePark: Street parking using mobile phones," in *Proceedings - 2012 IEEE 13th International Conference on Mobile Data Management, MDM 2012*, 2012, pp. 278–279. doi: 10.1109/MDM.2012.76.
[45] M. Balmer, R. Weibel, and H. Huang, "Value of incorporating geospatial information into the prediction of on-street parking occupancy–A case study," *Geo-Spatial Information Science*, vol. 24, no. 3, pp. 438–457, 2021, doi: 10.1080/10095020.2021.1937337.

[46] S. Yang and Z. (Sean) Qian, "Turning meter transactions data into occupancy and payment behavioral information for on-street parking," *Transportation Research Part C: Emerging Technologies*, vol. 78, pp. 165–182, 2017, doi: 10.1016/j.trc.2017.02.022.
[47] F. Bock, K. Xia, and M. Sester, "Mapping similarities in temporal parking occupancy behavior based on city-wide parking meter data," in *Proceedings of the International Cartographic Association (ICA)*, 2018, vol. 1, pp. 1–5. doi: 10.5194/ica-proc-1-12-2018.
[48] X. Ye, J. Wang, T. Wang, X. Yan, Q. Ye, and J. Chen, "Short-Term Prediction of Available Parking Space Based on Machine Learning Approaches," *IEEE Access*, vol. 8, pp. 174530–174541, 2020, doi: 10.1109/ACCESS.2020.3025589.
[49] R. F. Pozo, A. B. R. Gonzalez, M. R. Wilby, J. J. V. Diaz, and M. V. Matesanz, "Prediction of On-Street Parking Level of Service Based on Random Undersampling Decision Trees," *IEEE Transactions on Intelligent Transportation Systems*, 2021, doi: 10.1109/TITS.2021.3077985.
[50] Y. Liu, C. Liu, and X. Luo, "Spatiotemporal Deep-Learning Networks for Shared-Parking Demand Prediction," *Journal of Transprotation Engineering Part A - Systems*, vol. 147, no. 6, Jun. 2021, doi: 10.1061/JTEPBS.0000522.
[51] R. Domakuntla, "Identification of Parking Behaviour Changes using Car Fleet Data for On-Street Parking Information System," Technical University of Munich, 2021.
[52] J.-E. Navarro-B, M. Gebert, and R. Bielig, "On automatic extraction of on-street parking spaces using park-out events data," in *2021 IEEE International Conference on Omni-Layer Intelligent Systems (COINS)*, pp. 1–7.
[53] N. Bibi, M. N. Majid, and H. Dawood, "Automatic Parking Space Detection System," in *2017 2nd International Conference on Multimedia and Image Processing*, 2017, pp. 11–15. doi: 10.1109/ICMIP.2017.4.
[54] C. Ajeng and T. H. T. Gim, "Analyzing on-street parking duration and demand in a Metropolitan City of a developing country: A case study of Yogyakarta City, Indonesia," *Sustainability*, vol. 10, no. 3, 2018, doi: 10.3390/su10030591.
[55] S. Gutmann, C. Maget, M. Spangler, and K. Bogenberger, "Truck Parking Occupancy Prediction: XGBoost-LSTM Model Fusion," *Frontiers in Future Transportation*, vol. 2, no. July, pp. 1–17, 2021, doi: 10.3389/ffutr.2021.693708.
[56] N. Feng, F. Zhang, J. Lin, J. Zhai, and X. Du, "Statistical Analysis and Prediction of Parking Behavior," in *Network and Parallel Computing*, 2019. doi: 10.1007/978-3-030-30709-7_8.
[57] S. Gao, M. Li, Y. Liang, J. Marks, Y. Kang, and M. Li, "Predicting the spatiotemporal legality of on-street parking using open data and machine learning," *Ann GIS*, vol. 25, no. 4, pp. 299–312, 2019, doi: 10.1080/19475683.2019.1679882.

[58] C. Pflügler, T. Köhn, M. Schreieck, M. Wiesche, and H. Krcmar, "Predicting the Availability of Parking Spaces with Publicly Available Data," in *INFORMATIK 2016*, 2016, pp. 361–374.
[59] F. M. Awan, Y. Saleem, R. Minerva, and N. Crespi, "A Comparative Analysis of Machine / Deep Learning Models for Parking Space Availability Prediction," *Sensors*, vol. 20, no. 1, 2020, doi: https://doi.org/10.3390/s20010322.
[60] N. Arora *et al.*, "Hard to Park ? Estimating Parking Difficulty at Scale," in *The 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '19), August 4-8, 2019, Anchorage, AK, USA*, 2019, pp. 2296–2304.
[61] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, vol. 13-17-Augu, pp. 785–794. doi: 10.1145/2939672.2939785.
[62] W. Alajali, S. Wen, and W. Zhou, "On-street car parking prediction in smart city: A multi-source data analysis in sensor-cloud environment," in *Security, Privacy, and Anonymity in Computation, Communication, and Storage*, vol. 10658 LNCS, 2017, pp. 641–652. doi: 10.1007/978-3-319-72395-2_58.
[63] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, pp. 5–32, 2001, doi: doi.org/10.1023/A:1010933404324.
[64] C. Badii, P. Nesi, and I. Paoli, "Predicting Available Parking Slots on Critical and Regular Services by Exploiting a Range of Open Data," *IEEE Access*, vol. 6, pp. 44059–44071, 2018, doi: 10.1109/ACCESS.2018.2864157.
[65] Z. Zhao, Y. Zhang, and Y. Zhang, "A Comparative Study of Parking Occupancy Prediction Methods considering Parking Type and Parking Scale," *Journal of Advanced Transportation*, vol. 2020, 2020, doi: https://doi.org/10.1155/2020/5624586.
[66] A. Bhattacharyya, W. Wang, C. Tsang, and C. Amza, "Semantic-aware anomaly detection in real time parking data," in *2017 IEEE 15th International Conference on Industrial Informatics, INDIN 2017*, 2017, pp. 486–491. doi: 10.1109/INDIN.2017.8104820.
[67] C. Badii, P. Nesi, and I. Paoli, "Predicting Available Parking Slots on Critical and Regular Services by Exploiting a Range of Open Data," *IEEE Access*, vol. 6, pp. 44059–44071, 2018, doi: 10.1109/ACCESS.2018.2864157.
[68] B. Xu, O. Wolfson, J. Yang, L. Stenneth, P. S. Yu, and P. C. Nelson, "Real-time street parking availability estimation," in *Proceedings of the 2013 IEEE 14th International Conference on Mobile Data*, 2013, vol. 1, no. May 2016, pp. 16–25. doi: 10.1109/MDM.2013.12.
[69] S. J. Qin, "Data Science Education With Domain Knowledge and System Principles," *Harvard Data Science Review*, no. 3.2, pp. 1–7, 2021, doi: 10.1162/99608f92.6f037267.

[70] J. Heaton, "An Empirical Analysis of Feature Engineering for Predictive Modeling," in *SoutheastCon 2016*, 2016, pp. 1–6. doi: 10.1109/SECON.2016.7506650.

[71] D. Shipmon, J. Gurevitch, P. M. Piselli, and S. Edwards, "Time Series Anomaly Detection - Detection of Anomalous Drops with Limited Features and Sparse Examples in Noisy Highly Periodic Data," *Google, Inc.*, 2017.

[72] Y. Zheng, S. Rajasegarar, C. Leckie, and M. Palaniswami, "Smart Car Parking : Temporal Clustering and Anomaly Detection in Urban Car Parking," *2014 IEEE Ninth International Conference on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP)*, no. April, pp. 1–6, 2014, doi: 10.1109/ISSNIP.2014.6827618.

[73] Tomtom, "Maps & Map Updates," *Tomtom website*, 2021. https://www.tomtom.com/en_gb/sat-nav/maps-services/map-updates/#mydriveconnect-tab (accessed Aug. 27, 2021).

[74] S. Daskalov and V. Nikolov, "Prediction of Univariate Time Series Based on Clustering."

[75] Lean Startup Co. Education Program, "A Playbook for Achieving Product-Market Fit," *Lean Startup Co.* https://leanstartup.co/a-playbook-for-achieving-product-market-fit/ (accessed Feb. 24, 2022).

[76] Heap, "What is Product - Market Fit ? Creating A Data - Driven PMF," *Heap Website*. https://heap.io/topics/what-is-product-market-fit-creating-a-data-driven-pmf-framework (accessed Feb. 24, 2022).

[77] Microsoft, "Zoom levels and tile grid," 2020. https://docs.microsoft.com/en-us/azure/azure-maps/zoom-levels-and-tile-grid?tabs=csharp#quadkey-indices%5C (accessed Jun. 20, 2019).

[78] C. Dosch, M. Farghal, M. Kapsecker, C. Lorenz, and A. Mosharafa, "Parking Prediction in the City of Melbourne," 2017.

[79] I. S. Suwardi, D. Dharma, D. P. Satya, and D. P. Lestari, "Geohash index based spatial data model for corporate," in *The 5th International Conference on Electrical Engineering and Informatics: Bridging the Knowledge between Academic, Industry, and Community, ICEEI 2015*, 2015, pp. 478–483. doi: 10.1109/ICEEI.2015.7352548.

[80] G. Cantelmo, P. Vitello, B. Toader, C. Antoniou, and F. Viti, "Inferring Urban Mobility and Habits from User Location History," *Transportation Research Procedia*, vol. 47, pp. 283–290, 2020, doi: 10.1016/j.trpro.2020.03.100.

[81] L. Montini, N. Rieser-schüssler, A. Horni, and K. W. Axhausen, "Trip Purpose Identification from GPS Tracks," *Transportation Research Record: Journal of the Transportation Research Board*, no. 2405, pp. 16–23, 2014, doi: 10.3141/2405-03.

[82] D. Ettema, F. Bastin, J. Polak, and O. Ashiru, "Modelling the joint choice of activity timing and duration," *Transportation Research Part A: Policy and Practice*, vol. 41, pp. 827–841, 2007, doi: 10.1016/j.tra.2007.03.001.

[83] L. Gong, T. Morikawa, T. Yamamoto, and H. Sato, "Deriving Personal Trip Data from GPS Data : A Literature Review on the Existing Methodologies," *Procedia - Social and Behavioral Sciences*, vol. 138, no. 0, pp. 557–565, 2014, doi: 10.1016/j.sbspro.2014.07.239.

[84] M. Z. Rodriguez *et al.*, *Clustering algorithms : A comparative approach*. 2019.

[85] N. Piovesan, L. Turi, E. Toigo, B. Martinez, and M. Rossi, "Data Analytics for Smart Parking Applications," pp. 1–25, 2016, doi: 10.3390/s16101575.

[86] 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. doi: 10.1109/ICTAI.2014.49.

[87] J. H. Ward, "Hierarchical Grouping to Optimize an Objective Function," *J Am Stat Assoc*, vol. 58, no. 301, pp. 236–244, 1963.

[88] S. Fong, S. U. Rehman, K. Khan, K. Aziz, S. Sarasvady, and A. Vishwa, "DBSCAN : Past , Present and Future," in *The Fifth International Conference on the Applications of Digital Information and Web Technologies (ICADIWT 2014)*, 2014, pp. 232–238. doi: 10.1109/ICADIWT.2014.6814687.

[89] S. Singh, "Simple Random Sampling," in *Advanced Sampling Theory with Applications*, Springer, Dordrecht, 2003, pp. 71–136. doi: https://doi.org/10.1007/978-94-007-0789-4_2.

[90] C. Iliopoulou and K. Kepaptsoglou, "Combining ITS and optimization in public transportation planning: state of the art and future research paths," *European Transport Research Review*, vol. 11, no. 1, 2019, doi: 10.1186/s12544-019-0365-5.

[91] S. Anand, S. Gupta, V. Darbari, and S. Kohli, "Crack-pot: Autonomous Road Crack and Pothole Detection," *2018 International Conference on Digital Image Computing: Techniques and Applications, DICTA 2018*, pp. 9–14, 2019, doi: 10.1109/DICTA.2018.8615819.

[92] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python Fabian," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011, doi: 10.1289/EHP4713.

[93] K. Fawagreh, M. M. Gaber, and E. Elyan, "Random forests: from early developments to recent advancements," *Systems Science & Control Engineering : An Open Access Journal*, vol. 2, no. 1, pp. 602–609, 2014, doi: 10.1080/21642583.2014.956265.

[94] R. Moeckel and R. Donnelly, "Gradual rasterization: redefining spatial resolution in transport modelling," *Environment and Planning B: Planning and Design*, vol. 42, no. 5, pp. 888–903, Sep. 2015, doi: 10.1068/B130199P.

[95] R. Rothfeld, “Agent-based Modelling and Simulation of Urban Air Mobility Operation: An Evaluation of Travel Times and Transport Performance,” Technical University of Munich, 2021.

A Gomari et al. (2021). Cluster analysis of parking behaviour: A case study in Munich.

Reference: S. Gomari, C. Knoth, and C. Antoniou, “Cluster analysis of parking behaviour: A case study in Munich,” *Transp. Res. Procedia*, vol. 52, no. 2021, pp. 485–492, 2021. https://doi.org/10.1016/j.trpro.2021.01.057

The summary of this study can be found in Section 3.3.

Credit contribution statement:
The paper contributions per author are listed in bullets below:

- Syrus Gomari: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing-original draft, Writing-review & editing
- Christoph Knoth: Funding acquisition, Methodology, Project administration, Resources, Supervision, Writing-review & editing
- Constantinos Antoniou: Formal analysis, Methodology, Supervision, Writing-review

Available online at www.sciencedirect.com

ScienceDirect

Transportation Research Procedia 52 (2021) 485–492

Transportation Research Procedia

www.elsevier.com/locate/procedia

23rd EURO Working Group on Transportation Meeting, EWGT 2020, 16-18 September 2020, Paphos, Cyprus

Cluster analysis of parking behaviour: A case study in Munich

Syrus Gomari[a,b,*], Christoph Knoth[a], Constantinos Antoniou[b]

[a]BMW Group, Max-Diamand-Str. 25, Munich 80937, Germany
[b]Technical University of Munich, Arcisstr. 21, Munich 80333, Germany

Abstract

Estimates show that vehicles cruising for on-street parking contribute to 30% of urban traffic congestion. On-street parking information (OSPI) systems are increasingly becoming a more popular service to help lessen the on-street parking search time and consequently reduce congestion. However, despite the service offerings of these prediction models, the on-street parking behaviour of people in cities have not been studied to the same magnitude. The lack of appropriate empirical parking data is one main reason. This study focuses on the analysis of parking behaviour by capturing the on-street parking dynamics, which can give a better insight on a city's parking contextualization. The case study examined is the parking behaviour dynamics within Munich by inferring from parked-in and parked-out events data from vehicles. A two part clustering analysis was conducted: (1) agglomerative clustering on the temporal trend of parking dynamics (TTPD) and (2) a two-stage DBSCAN – K-means clustering on the parking duration information. The results show that using the methodology introduced, the parking behaviour within the city can be obtained using this unsupervised learning approach.

Keywords: Smart parking; vehicle parking events; parking behaviour; clustering; Munich, Germany

1. Introduction

Vehicles cruising for parking contributes to substantial congestion within an urban transport network (Friedrich et al., 2019). As a parking management measure, parking guidance signs have been placed within a transport network to guide vehicles to predominantly off-street parking options. Comparable systems have also recently been developed for finding parking spots on the streets, denoted as on-street parking information (OSPI). State-of-the-art OSPI systems are mostly developed using complex machine learning techniques aiming to optimise prediction estimates without necessarily investigating the underlying parking behaviour in a city. The quality of the information provided by such systems are validated by the comparison of observed on-site data against the prediction model estimates. Although many forms of ground truth (GT) strategies exist, there is still no scalable method that can significantly cut

* Corresponding author. Tel.: +49 151 601 36386.
E-mail address: syrus.gomari@bmw.de

10.1016/j.trpro.2021.01.057

down data collection costs. The hypothesis being tested in this study is that clustering the parking events data can give us a better insight about the parking dynamics within a city. The results can then give guidance to do targeted, as opposed to random, ground truth collection to truly validate the models on all important facets, and not have an imbalanced training data. The parking behaviour dynamics is analysed within Munich, Germany by inferring from a new type of dataset, which is parked-in and parked-out events data generated from a fleet of vehicles. This paper primarily focuses on the clustering of temporal dynamics of the parking events and the parking duration. The paper is further structured by first a review of studies, in particular, the type of data used for the parking information researches, followed by the description of the parking events dataset that is used in this study. A methodology for the cluster analysis to determine the parking behaviour is then elaborated. The results of the clusters generated are examined in the analysis section. The paper ends with conclusions that can be drawn and with recommended applications.

2. Literature Review

There have been studies that have tackled the challenges of estimating mobility behaviour and approximating the possible trip purpose or activity done by a group of users inferring from GPS data. (Cantelmo et al., 2020; Ettema et al., 2007; Gong et al., 2014; Montini et al., 2014) However, there have not been many studies that tackle this in the area of parking behaviour. The majority of studies in the area of parking have been on prediction models, where parking behaviour or model interpretation is not the focus, but accuracy. The models use a diverse range of data sources to train, validate, and test their complex machine learning models. The datasets could potentially have also been used for understanding parking behaviour despite being spatially and temporally limited. Relevant for this paper is to get an overview of the ground truth *parking* data used for within their research and qualitatively compare them with the parking events dataset used in this study.

Smart parking meters is one of the data sources researchers have been looking into as mentioned in Bock & Di Martino (2017) and Yang & Qian (2019). Liu et al. (2018), Shao et al. (2018), and Monteiro & Ioannou (2018) developed on-street parking guidance systems using data gathered from on-street parking spaces with sensors. Gkolias & Vlahogianni (2018) obtained parking data from fewer than 10000 images captured by a camera on a moving vehicle. All of these researches were successful in creating a prediction model, despite the spatial limitation in their data collection. The parking events dataset used in our study has an advantage over these studies spatially distributed without limitation, in spite of fleet drivers' behaviour constraints. Furthermore, the parking events data is continuously gathered and increasing in volume as more vehicles are equipped.

3. Data and Study Area

3.1. Study area and data description

The parking events dataset in this study contains geolocation and temporal information of parked positions of BMW vehicles in Munich for the month of July 2019. The data collection is done by BMW's backend services including filtering steps to ensure complete anonymisation according to EU defined data privacy standards. A parking event is generated after a certain minimum time threshold since the vehicle has switched on or off the engine, indicating a parked-out and a parked-in position, respectively. The anonymised data collected contains only parking events within the proximity of a street. This limits the study to BMW's 2019 OSPI service area for the city of Munich, Germany (see Figure 1a). For this study, the sample taken contains no mid-week public holidays and it is assumed to be representative of regular weekday and weekend conditions. The number of parking events even gathered for just a day is more than the manual observations made by on-site surveyors over a year. This indicates the potential of extracting the parking dynamics situation from a data source that is more accessible and widespread spatially and temporally. The parking events data can be paired within the defined service area (see Figure 1); green dots are parked-out events while red dots are parked-in. This is significant in terms of the usage of the data and allows the extraction of duration information. On an overall level (see Figure 1c), the mean duration of the dataset is at 272 minutes with a median of 48 minutes, indicating right skewness. The parking events data only up to the 95th percentile was taken with the value at 1069 minutes; the remaining 5th of a percentile stretches the dataset's maximum duration to 32774 minutes. This preprocessing step was done to cut out large outliers from the dataset.

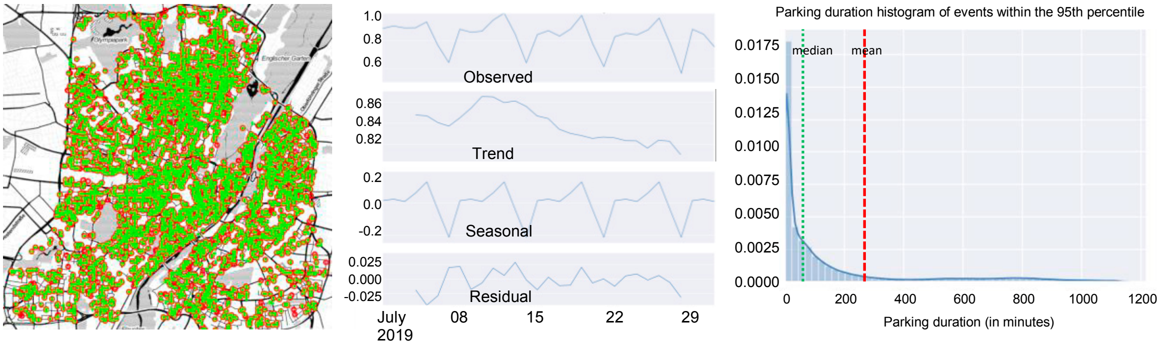


Figure 1. Paired parking events for one day (left); b) time series decomposition of parked-in events (centre); c) parking duration histogram (right)

Mobility data is normally assumed as a time series data that follows a trend. The additive time series decomposition unravels the different trends present in parking events dataset for a month. Figure 1b shows the slight overall decrease in the observed volume towards the end of July. This can be attributed to inhabitants of Munich going on holidays. Specifically, looking at the trend graph, for parked in events there is a decline, which can indicate a trend of more vehicles leaving Munich towards the end the month than coming in. The seasonal graph illustrates a weekly pattern and the residual graph shows the random trends that occur constituting a maximum of ±3% randomness.

4. Methodology

Clustering, an unsupervised learning approach, was the chosen method for identifying on-street parking event clusters. This method captures and partitions similar patterns that are difficult to identify manually (Zheng et al., 2014). The parking behaviour cluster analysis is divided into two parts (Figure 2): (1) parking dynamics: the grouping of quadkey-level districts by the temporal trend of parking dynamics (TTPD) and (2) parking purpose by parking duration. A quadkey[1] is an indexing naming convention and unique identifier of a standard map tile on a particular zoom level. This standardized division of the world map into tiles is a standard used by Microsoft's Azure Maps.

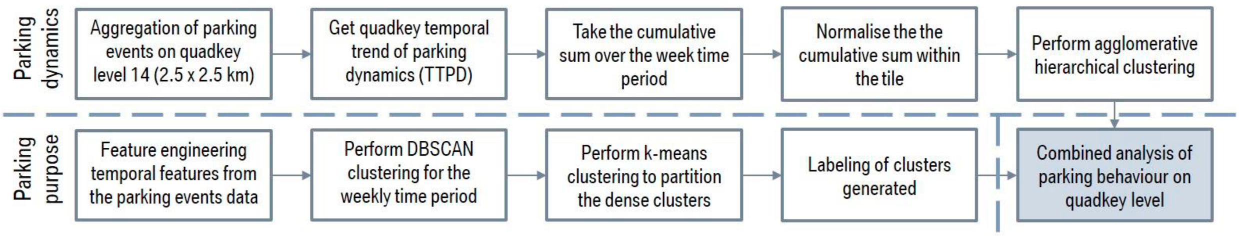


Figure 2. Workflow of the methodology

4.1. Parking dynamics: Clustering of quadkey-aggregated parking events based on TTPD

The raw parking events attributes used in the study were parking event type (parked-in or parked-out), timestamp, latitude, and longitude. The first data processing step was to aggregate parking events by quadkey. The quadkey zoom level 14 (2457.6 x 2457.6 m) was the selected optimal tile size to reduce relative error propagation due to low aggregated volume since the data is only for one month. A finer tile level would reduce the number of parking events per tile, and thereby increasing relative error. The quadkey approach is favorable to generate reproducible and comparable results for the same application in future researches. The cumulative sum, at each quadkey, of the net parking is calculated by the summation of the difference between parked-in (**PIN)** and parked-out (**POUT)** events per

[1] See definition of quadkey at: https://docs.microsoft.com/en-us/azure/azure-maps/zoom-levels-and-tile-grid?tabs=csharp#quadkey-indices

15-minute intervals aggregated to a weekly period. This is to estimate the activity of parking happening in each quadkey, referred to hereafter as temporal trend of parking dynamics (**TTPD**) defined by the time-series:

$$TTPD_q^T = \sum_{t=0}^{T} PIN_q^T - POUT_q^T \tag{1}$$

where parked-in ($\mathbf{PIN_q^T}$) and parked-out ($\mathbf{POUT_q^T}$) events are time series vectors of 15-minute aggregated parking events at each quadkey, q (i.e. $\boldsymbol{P_q^T} = \{P_{qt}^T; q = 1, 2, \ldots, N; t = 00{:}00, 00{:}15, 00{:}30, \ldots, T\}$, for N number of quadkeys, and length of study time period, T, which is one week. The time series are normalised to get the relative values for comparability of quadkeys. The final step was to cluster the time series on quadkey-level by agglomerative hierarchical clustering using Ward's algorithm (Ward, 1963). This hierarchical clustering method is widely used and has been popular for its interpretability through a dendrogram and step-wise approach of starting with all objects as one cluster and joining, at each step, the two most similar clusters. Essentially, no input is required, with number of clusters being defined afterwards. The silhouette score metric (Rousseeuw, 1987)was also used as a guide to select the number of clusters. The output clusters are groups of the most similar quadkey-level parking dynamics time series over the study period.

4.2. Parking purpose: Clustering of individual parking events based on parked-in time and parking duration

The parking purpose was categorised using parking duration and parked-in time. The parked-in time was converted to time during the day in seconds and analysis was done on a week unit as the time period. The parking duration (*PD*) was calculated by matching the Geohash (i.e. geocode for specific pair coordinates) of parked-in and parked-out events and taking their difference; in this case each pair match can essentially be treated as a remote on-street parking sensor:

$$PD = POUT_{timestamp} - PIN_{timestamp} \tag{2}$$

Parked-in time is the start time of an activity or walk to a driver's destination, while parking duration indicates the time spent in a neighbourhood. These two features were inputs to the two-staged clustering using density-based spatial clustering of applications with noise (DBSCAN) and K-means. DBSCAN was first applied to cluster according to the densities formed by similar observations and to recognise if there are distinct groups that are separate according to their density cluster distances from one another. The method was selected as it has been widely recommended for its ability to identify shape patterns in large datasets. It was discovered from the exploratory data analysis that the majority of observations were close and densely packed in one whole cluster regardless of the features used. Hence, partitioning-based clustering was needed by K-means to further create further clusters. This approach was taken to automate the partitioning of the parked-in time and duration instead of slicing the unlabelled data manually.

5. Cluster Analysis

To capture the on-street parking behaviour, the analysis was done jointly on parking dynamics and parking purpose. First, quadkeys were identified that could be similar in terms of when they fill up and when they empty using the TTPD; second, clusters were identified from the parked-in time and parking duration. Then, the analysis is done by interpreting the distribution share of the parking purpose clusters in each quadkey category cluster.

5.1. Parking dynamics clusters based on TTPD

The spatial spread of parking events within Munich is mostly concentrated on a few quadkeys. Figure 3 shows that the top 2 out of the 23 quadkeys contain about 25% of all events, the top 5 comprises more than 50%, while the bottom 12 quadkeys encompass only 25% of all parking events. The majority of the events take place around the city centre and a secondary transport hub to north of the Munich polygon (see Figure 3). From this perspective, the relative error is large for quadkeys with lower volume of parking events, despite potentially behaving the same as quadkeys with a large volume. The low volume is interpreted in two ways: one is that there are no BMW vehicles present in these areas or there are parking events occurring but not captured by the dataset. Since Munich has a good coverage of BMW, it is assumed that the dataset is representative of the population.

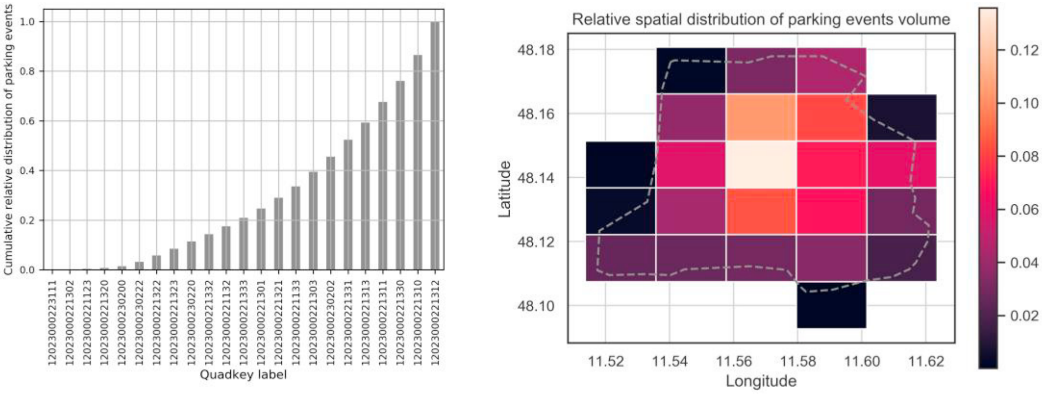


Figure 3. Cumulative distribution of parking events by quadkey (left) and its relative spatial distribution (right)

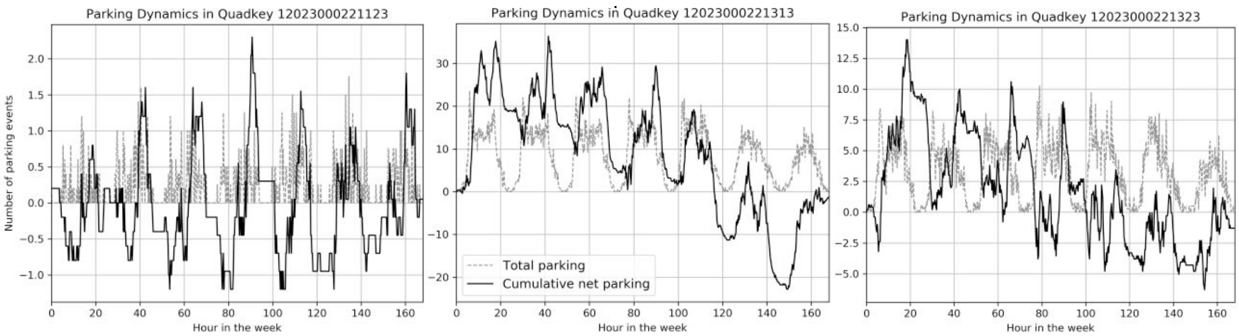


Figure 4. a) Quadkey categories 0: Eating, but low volume of events (left); 1: b) Business and shopping (center); 2: c) Business and eating (right)

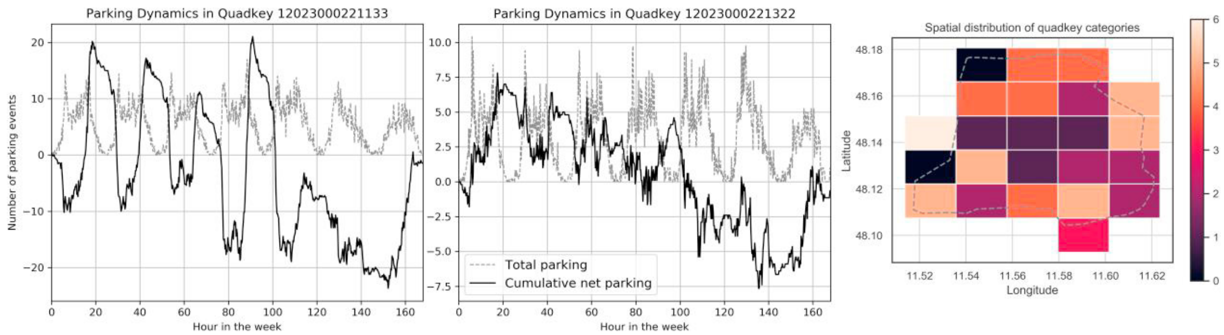


Figure 5. a) Quadkey categories 4: Residential (left); b) 5: Residential, shopping, and eating (middle); c) Spatial distribution of categories (right)

The TTPD for each quadkey was calculated and agglomerative hierarchical clustering (see Section 4.1) was performed. Seven clusters were identified based on the silhouette score metric, the dendrogram, and manual labeling of the 23 quadkeys based on their TTPD and other parking dynamics indicators, such as, total parking (Figure 4b) , net parking, and the difference of the two. Only 5 out of the 7 clusters were significant, the other 2 had low volumes of parking events (i.e. categories 3 and 6). The manual labeling was done to compare and validate the robustness of the clustering technique, which turned out to be confirmatory. An example for each identified quadkey category is illustrated in Figure 4 and Figure 5. The cumulative net parking and the total parking are displayed in these figures, which represent the rate of filling in of a quadkey and the volume of parking events at each time step, respectively.

Category 1 ‘eating’ cluster (Figure 4a) shows activity at lunch and dinner hours; category 2 ‘business and shopping’ (Figure 4b) covers the usual working hours with some activity going in between the peak hours; category 3 ‘business and eating’ (Figure 4c) shows a TTPD peaking parked-in events in the morning rush hour, lunch time, afternoon rush hour and continuing until dinner time and weekend dining patterns; while category 4 ‘the residential’ cluster (Figure 5a) shows the filling in of the quadkeys at night time without any activity until the morning peak hour; and category 5 (Figure 5b) is a mixture of ‘residential, shopping, and eating’, where some parking activities is observed during the day and peaking towards the dinner hours and stabilizing until next day’s morning peak hour and with no day shopping activities on Sunday but with a peak in the evening. The clustering with TTPD on 15-minute time intervals shows that meaningful clusters for parking dynamics can be extracted on a high level. The next 1 aims to go further into detail with regards to understanding the purpose for parking in the quadkeys based on duration.

5.2. Parking purpose clusters based on parked-in time and duration

The input for the parking purpose clustering was mainly the parked-in time of a vehicle and the duration of stay in the area. The entire dataset was fed into the clustering algorithm without prior geographical grouping. The identified clusters of parking purpose in combination with the parking dynamics by quadkey clusters (see Section 5.1) gives an estimate of the parking behaviour as a whole in each quadkey. Based on initial exploratory data cluster analysis, the hypothesis of different parking behaviours during weekdays and weekends is valid and separation improves clustering results. Hence, two temporal grouping were introduced: weekdays and weekends. Different separations of days and time periods were tested, and the most logical result was to create a separation for Monday to Friday evening, and Friday evening to Sunday.

The results of the first clustering step using DBSCAN are shown in Figure 6. There is a distinction between two large clusters during the entire period. These are shorter term parking during the day and longer term overnight parking towards the night. The noise from the results of the DBSCAN clustering and the night clusters for both temporal groupings are left out for the next step. The night cluster though is labeled as overnight parking purpose cluster and included in the final cluster output. The results of the K-means clustering are shown in Figure 6. There are 6 clusters for the weekday grouping, and 9 clusters for the weekend. The former has lower number of clusters, since there are more regular daily activities on weekdays, as opposed to random activities on weekends. Overall, the two-staged clustering aids in identifying more clusters, compared to doing it using only DBSCAN.

The important details of the 16 parking duration purpose clusters are summarised in Table 1. The summary includes a brief cluster description, descriptive statistics about the duration in minutes and the parked-in hour, the volume share of parking events per cluster, and the spread of these parking events across the 5 quadkey categories. A heat map color map is applied to illustrate the difference in magnitude of the events share in each quadkey category at each parking duration purpose cluster. Each table cell in the heat map describes the parking event joint probability on that cell.

The share of parking events by quadkey is 37.1%, 24.0%, 22.3%, 16%, and 0.7% for category 2, 4, 3, 5, and 1, respectively. The change in order of share within a certain purpose cluster can describe its parking behaviour. The high-share parking event clusters can be seen on weekdays. The top 3 highest, comprising of about 19% of all events, occur on quadkey category 2 (business and shopping) at parking purpose clusters 1, 4, and 6, which are early morning short-term parking (mean parked-in time of 7:00), lunch time parking (12:00), evening peak (17:00) hour short-term parking, correspondingly. The average parking duration in these 3 clusters is circa 30 minutes. The share order is not disrupted much, indicating stability on weekdays. The order change only happens on purpose cluster 5, where quadkey category 3 comes in second highest share. Cluster 5 overlaps with lunch time and includes intermediate parking to do other activities other than going home (category 4) until after dinner time with a mean duration of 180 minutes. On weekends, the peaks occur from 10:00 to 15:00 and 17:00 to 20:00. Five out of nine clusters, these are clusters 9, 10, 12, 14, and 15 have a higher share on quadkey category 3 (business and eating) than 4 (residential). These happen to be overlapping with the two peak hours, thus, the reason is most likely eating or longer term shopping.

Another cluster which was separated after the DBSCAN clustering process, which was labeled as overnight parking purpose proved to be correct. As it can be seen, there is a high share of long-term overnight parking occurring at quadkey category 4 and 5 (relative to its overall share), which were interpreted and labeled as residential from the TTPD graphs (Figure 5).

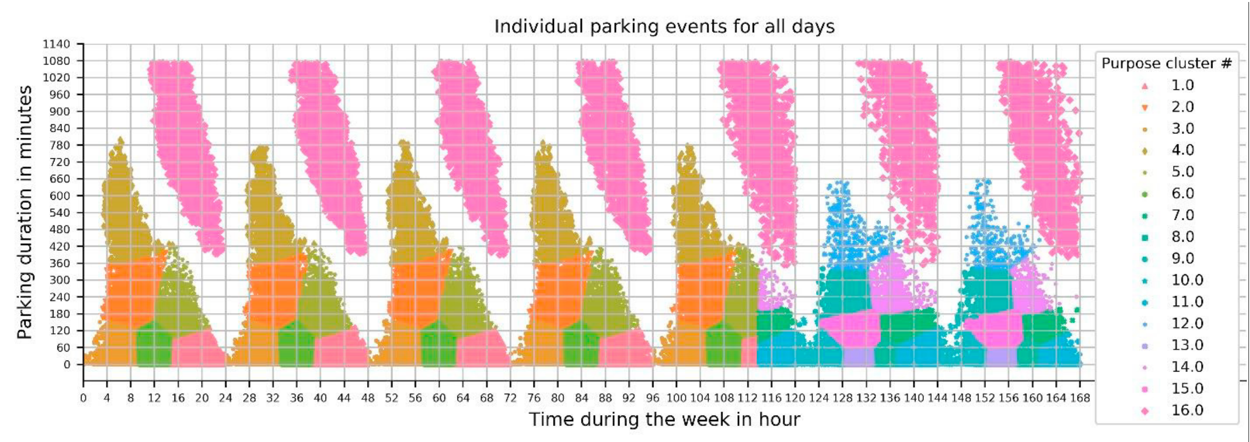


Figure 6. Generated parking purpose (duration) clusters using DBSCAN and K-means clustering

Table 1. Summary of the details for each parking purpose duration cluster

Purpose cluster number	Temporal Grouping	Purpose cluster description	Parking duration (min.)			Parked-in time (hour)			Mean parking events share	Quadkey category share				
			Mean	Min	Max	Mean	Min	Max		0	1	2	4	5
1	Weekday	Early morning short parking	30	0	155	6.90	0.00	9.42	18.2%	0.11%	6.27%	4.06%	4.45%	3.33%
2		Morning peak hour parking to work	511	358	798	7.35	3.32	15.01	2.5%	0.01%	1.04%	0.44%	0.48%	0.56%
3		Intermediate parking before lunch	228	120	401	8.78	3.22	13.89	4.7%	0.02%	1.87%	0.93%	0.95%	0.89%
4		Lunch time parking	34	0	150	11.89	9.27	14.58	20.4%	0.12%	7.24%	4.50%	4.93%	3.66%
5		Intermediate parking after lunch	182	94	425	15.92	11.84	21.56	7.2%	0.06%	2.95%	1.69%	1.50%	1.01%
6		Evening peak short parking	27	0	132	17.02	14.40	23.98	16.6%	0.10%	5.82%	3.98%	4.06%	2.70%
7	Weekend	Early morning short parking	20	0	162	6.66	0.00	8.74	2.6%	0.01%	0.86%	0.58%	0.68%	0.46%
8		Long-term parking during the day	440	339	658	9.28	4.74	17.81	0.4%	0.00%	0.16%	0.09%	0.12%	0.06%
9		Morning parking for shopping and/or dining	234	155	340	9.45	3.37	13.05	1.0%	0.01%	0.41%	0.23%	0.22%	0.15%
10		Morning parking for weekly shopping/dining	112	65	188	10.62	4.53	13.94	2.4%	0.02%	0.98%	0.58%	0.51%	0.33%
11		Morning short parking	18	0	67	10.75	8.64	12.93	4.2%	0.03%	1.38%	1.01%	1.11%	0.68%
12		Afternoon short parking	20	0	90	15.03	12.81	17.48	4.1%	0.02%	1.45%	1.01%	0.99%	0.60%
13		Afternoon shopping and/or dining	252	182	403	16.21	12.46	23.95	1.0%	0.01%	0.42%	0.20%	0.22%	0.15%
14		Evening dining/pubs	123	62	202	17.00	13.77	23.93	2.0%	0.02%	0.87%	0.50%	0.42%	0.23%
15		Evening short parking	18	0	125	19.87	17.44	24.00	2.9%	0.02%	1.13%	0.74%	0.70%	0.35%
16	Overnight	Overnight long-term parking at residence	770	355	1075	17.59	10.40	23.96	9.6%	0.05%	2.19%	1.85%	3.47%	2.01%
						Parking events share in each quadkey:				0.7%	37.1%	22.3%	24.0%	16.0%

6. Conclusions and recommendations

The paper has given insights on estimating on-street parking behaviour from BMW's parking events data. The analysis was done on two parts: the temporal trend of parking dynamics (TTPD) and the parking purpose based on duration. The first aspect was formulated based on the aggregation of the parking events on a quadkey. Quadkey zoom level 14 was selected as the optimal trade-off between size of the tile and relative noise based on the volume of parking events per quadkey. The volume of parking events in each quadkey was feature engineered into a weekly time series. This entails getting the cumulative distribution of the difference between the parked-in and parked-out events on 15-minute intervals. This transformation was used as an indicator to detect if an area is filling up or emptying. Then, the quadkeys were clustered and categorised using agglomerative hierarchical clustering. The second aspect was identifying the parking purpose based on the attributes parked-in time and the parking duration. Given the nature and spread of the data, a two-step clustering approach using DBSCAN and K-means was applied to partition the parking events. Overlapping the results of the two parts enabled the better capturing of the overall parking behaviour. The outcome gave us the joint probability of parking purpose by duration and quadkey category. This collectively generated the general parking behaviour insights within a study area using the parking events dataset.

The study area in this paper covers a dense urban area close to Munich's city centre with a mixture of commercial and residential landuse. This creates a bias towards short term parking behaviour considering the predominantly dense parking situation. Through the introduced unsupervised learning methodology, we were able to further categorise areas by quadkeys into a combination of residential, business, eating, and shopping areas. However, a further categorisation of short-term parking activities, especially shopping, is difficult to identify without driver input about trip purpose. To further go into detail, a recommended approach is to use the parking behaviour results in combination with inferring from cleaned OpenStreetMap points of interest (POIs) data. This helps better comprehend possible correlations between quadkeys and an activity performed resulting to more localised estimations and the detection of popular areas in time and space.

Automating processes, especially of data collection, has become quite popular. Another application of the results of this study is leveraging the value of an automated data collection system such as parking events. The ability to infer parking behaviour from the inflow of parking events data automatically could support testing and validation efforts for parking-related products and research. The findings from this paper are a guiding step to enable the better measurement of the quality of prediction models, given that the parking events data is widespread spatially and temporally, as opposed to manual data collection.

The methodology applied in this paper can be extended to all other cities where parking events data is being gathered. The method can immediately provide first insights on the spatio-temporal parking behaviour that exists within a city while employing a random automated data collection by a fleet of vehicles representing normal human mobility behaviour, with a bias towards the group of vehicle users. This study will directly feed into a bigger research on ground truth strategies for parking prediction by applying the method to all the available data for a longer period.

7. References

Bock, F., Di Martino, S., 2017. How many probe vehicles do we need to collect on-street parking information? 5th IEEE Int. Conf. Model. Technol. Intell. Transp. Syst. MT-ITS 2017 - Proc. 538–543. https://doi.org/10.1109/MTITS.2017.8005731

Cantelmo, G., Vitello, P., Toader, B., Antoniou, C., Viti, F., 2020. Inferring Urban Mobility and Habits from User Location History. Transp. Res. Procedia 47, 283–290. https://doi.org/10.1016/j.trpro.2020.03.100

Ettema, D., Bastin, F., Polak, J., Ashiru, O., 2007. Modelling the joint choice of activity timing and duration. Transp. Res. Part A Policy Pract. 41, 827–841. https://doi.org/10.1016/j.tra.2007.03.001

Friedrich, T., Krejca, M.S., Rothenberger, R., Arndt, T., Hafner, D., Kellermeier, T., Krogmann, S., Razmjou, A., 2019. Routing for On-Street Parking Search using Probabilistic Data. AI Commun. Pre-press, 1–12.

Gkolias, K., Vlahogianni, E.I., 2018. Convolutional Neural Networks for On-Street Parking Space Detection in Urban Networks. IEEE Trans. Intell. Transp. Syst. PP, 1–10. https://doi.org/10.1109/TITS.2018.2882439

Gong, L., Morikawa, T., Yamamoto, T., Sato, H., 2014. Deriving Personal Trip Data from GPS Data : A Literature Review on the Existing Methodologies. Procedia - Soc. Behav. Sci. 138, 557–565. https://doi.org/10.1016/j.sbspro.2014.07.239

Liu, K.S., Gao, J., Wu, X., Lin, S., 2018. On-Street Parking Guidance with Real-Time Sensing Data for Smart Cities, in: 15th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON). IEEE, Hong Kong, China. https://doi.org/10.1109/SAHCN.2018.8397113

Monteiro, F. V, Ioannou, P., 2018. On-Street Parking Prediction Using Real-Time Data, in: 2018 IEEE 21th International Conference on Intelligent Transportation Systems (ITSC). IEEE, Maui, Hawaii, USA, pp. 2478–2483.

Montini, L., Rieser-schüssler, N., Horni, A., Axhausen, K.W., 2014. Trip Purpose Identification from GPS Tracks. Transp. Res. Rec. J. Transp. Res. Board 16–23. https://doi.org/10.3141/2405-03

Rousseeuw, P.J., 1987. Silhouettes : a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. 20, 53–65.

Shao, W., Zhang, Y., Guo, B., Qin, K., Chan, J., Flora, D., 2018. Parking Availability Prediction with Long Short Term Memory Model. GPC 2018 1–15.

Ward, J.H., 1963. Hierarchical Grouping to Optimize an Objective Function. J. Am. Stat. Assoc. 58, 236–244.

Yang, S., Qian, E., 2019. A deep learning approach to real-time parking occupancy prediction in spatio-temporal networks incorporating multiple spatio-temporal data sources Uses parking meter transactions, traffic conditions, weather conditions. Transp. Res. Part C.

Zheng, Y., Rajasegarar, S., Leckie, C., Palaniswami, M., 2014. Smart Car Parking : Temporal Clustering and Anomaly Detection in Urban Car Parking. 2014 IEEE Ninth Int. Conf. Intell. Sensors, Sens. Networks Inf. Process. 1–6. https://doi.org/10.1109/ISSNIP.2014.6827618

B Gomari et al. (2022). Prioritization-based subsampling quality assessment methodology for mobility-related information systems

Reference: S. Gomari, C. Knoth, and C. Antoniou, "Prioritization-based subsampling quality assessment methodology for mobility-related information systems," *IET Intell. Transp. Syst.*, vol. 16, no. 5, pp. 602-615, 2022. https://doi.org/10.1049/itr2.12160

The summary of this study can be found in Section 3.4.

Credit contribution statement:
The paper contributions per author are listed in bullets below:

- Syrus Gomari: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing-original draft, Writing-review & editing
- Christoph Knoth: Funding acquisition, Methodology, Project administration, Resources, Supervision, Writing-review & editing
- Constantinos Antoniou: Formal analysis, Methodology, Supervision, Writing-review

Received: 10 May 2021 | Revised: 11 November 2021 | Accepted: 8 December 2021

IET Intelligent Transport Systems

DOI: 10.1049/itr2.12160

The Institution of Engineering and Technology WILEY

ORIGINAL RESEARCH PAPER

Prioritization-based subsampling quality assessment methodology for mobility-related information systems

Syrus Gomari[1,2] | Christoph Knoth[3] | Constantinos Antoniou[1]

[1] Chair of Transportation Systems Engineering, Technical University of Munich, Munich, Germany

[2] Connected Parking Team, BMW Group, Munich, Germany

[3] Analog and RF Verification, Infineon Technologies AG, Munich, Germany

Correspondence
Syrus Gomari, Chair of Transportation Systems Engineering, Technical University of Munich, Arcisstr. 21, Munich 80333, Germany.
Email: syrus.gomari@tum.de; syrus.gomari@bmw.de; syrus.gomari@gmail.com

Abstract

Mobility-related information systems, such as on-street parking information (OSPI) systems have become more popular in the original equipment manufacturer (OEM) industry over the last decade. However, there is a lack of methods to assess their quality at a large scale. This paper introduces a data-driven methodology to measure the true quality by fleet data prioritization-based subsampling strategies (PSSs). It is applied to the use case of OSPI using parking events (PE), but is applicable to other mobility-related information systems utilizing their respective fleet data. PSSs are defined based on neighbourhoods and time periods. Each PSS generates a unique set of spatio-temporally important areas at different quadkey zoom levels over 168 week-hours, called slices. The importance weight in each slice depends on the volume of PE within them. The algorithm for each PSS automatically selects important areas and time frames that are vital to be observed. Sample prediction models are used for the benefits assessment of the methodology by comparing it against non-prioritized randomized selection of ground truth. It is proven that the methodology can lessen the effort of ground truth collection, while maintaining the amount of information necessary to assess the true quality of a prediction model.

1 | INTRODUCTION

1.1 | Background on quality assessment of mobility-related information systems

Quality assessment (QA) of mobility-related information systems (IS) has mainly focused on measuring the discrepancies in the technical broadcasting and availability of information [1]. The assessments do not necessarily evaluate the accuracy of the information's content [1]. Existing QA in the area of mobility, do not consider the relative importance of information given to users. For example, the importance of correctly relying information to a user about a train with a 15-min headway is higher than a train that arrives every 2 min. Another instance is, information about vacant on-street parking is more important for a driver in a busy central area compared to parking availability in the periphery of a city with minimal traffic. [2] and [3] refer to this as the gap between the delivered information quality by a service provider and the users' expected quality based on perceived utility. The quality of an IS needs to be assessed based on the features important to the system objectives and user or management expectations [4]. Essentially, to assess the true quality of an IS, the evaluator must comprehend the needs of its users and satisfy them to the highest quality. Although quality assessment methods exist in mobility-related information systems, to the best knowledge of the authors, there is a gap in knowledge for comprehensive prioritization-based methods. To address this gap, in this paper, a methodology is introduced that describes a procedure on utilisation of fleet data for defining prioritization-based subsampling for quality assessment. Furthermore, the viability of the method is demonstrated by assessing the quality of on-street parking information (OSPI) systems delivered by different prediction models. OSPI is a chosen special case where higher efforts are required for QA in comparison to traffic for instance. OSPI involves a high number of small streets where low volume of on-street parking occurs, whereas the traffic deals with observing a low number of major roads where high volume traffic is easier measured. This makes OSPI QA comparably more error-prone, and thus, higher efforts and more precise QA methods are needed. As a limitation in this research, software and system quality are not tackled and are out of scope.

1.2 | Use case background: On-street parking information (OSPI)

Vehicles cruising for on-street parking contribute to a significant amount of congestion within a city's inner urban area [5, 6]. Based on 22 studies in different cities ranging from 1927 to 2015 as discussed in [6], the average cruising traffic share in a city is around 34% and drivers spent around 8 min searching for parking. OSPI services exist as a guidance system to smartly navigate drivers in search for on-street parking. A couple of benefits of OSPI are the reduction of traffic congestion caused by cruising drivers [7–10] and pre-departure information of parking situation at destination that increases the chances of finding a parking spot [11]. The latter can even help drivers decide whether it is wise to take their vehicles. The state-of-the-art OSPI systems are mostly developed using complex machine learning techniques [7, 8, 10, 12–18]. The majority of models aim to achieve real-time prediction, but there has also been a study on estimating parking availability for a given time interval, like 10–20 min [19]. Despite advances in artificial intelligence, OSPI services still have yet to entice the majority of potential users, and hence, there is still potential to attract more users to increase benefits on a system level. Further added value for drivers comes with the capability to correctly assess the quality of a service. Thus, as an initial step, the true quality of OSPI needs to be assessed, which entails considering the relative importance of the information delivered to drivers, and thereby satisfy their needs. After all, the true quality of such systems determines the benefits gained in a transport network. True quality in this paper refers to the adjusted quality metric scores based on important or prioritized areas (see Section 2).

The main difference between state-of-the-art OSPI models available and how they are validated is the data gathered and the features considered for training, validating, and testing the models [19]. Data sources that have been used to validate parking prediction models are: smart parking meters [15, 18, 20, 21], mobile payments [8, 22, 23], intelligent parking systems [24], real-time ground sensors [14, 17, 25, 26] images captured by a camera mounted on a moving vehicle [7, 27], crowd-sensing information by equipping probe vehicles (e.g. taxis) with on-board sensors, cameras, or ultrasonic sensors [28, 29], or crowd-sensing using GPS signals from smartphones [23, 29–31], and also manual observations [32]. A study aiming to improve automatic extraction of parking spaces used on-street parked out events from connected vehicles to validate legal and illegal parking spaces in the city [33]. The differences in input data play a major role in the reliability and quality. The information quality of models in the studies was validated by the comparison of randomly observed ground truth (GT) data against prediction availability estimates.

1.3 | Significance of prioritization-based subsampling for quality assessment

Although many forms of GT strategies exist, there is still no scalable method that can reduce data collection efforts and costs. Some alternatives are to randomly reduce subsamples, which is tested in this study (see Section 3.4), or acquire local knowledge about the landuse and daily parking behaviour. However, since these methods are labour-intensive, they are not scalable. Thus, a fully automated prioritization method is sought to reduce ground truth efforts and thereby reduce costs, while maintaining and also potentially improving the system.

The hypothesis tested in this study is that with a data-driven methodology using fleet data; it is possible to get a better insight for targeted and prioritization-based subsampling GT collection strategies. No studies exist that provide a prioritized-based subsampling of GT for quality assessment since most are based on fixed sensors or parking meters and lack large amounts of data to prioritize areas. This paper looks into the potential usage of vehicle parking events as a source for prioritizing ground truth collection at neighbourhoods, which are selected based on the frequency of visits within a certain time bucket, called slices. Identifying such priority slices assist GT collection efforts in areas which are important for customers to have relevant accurate dynamic parking information. Developing a methodology considering strategical slices of a GT collection set gives a complete picture of the service quality.

The main contribution of this study is the development of a methodology that measures the true quality of competitive mobility-related prediction models (see Section 2) and can provide recommendations to reduce the required ground truth data for quality assessment. The true quality is assessed by assigning importance weights to areas and time periods based on the chosen fleet volume (e.g. parking events, traffic flows). The methodology is applied on the use case of on-street parking (see Section 3). The main findings are described in Section 3.4, and a summary of contributions are described in the last section.

2 | METHODOLOGY: USING VEHICLE FLEET DATA FOR QUALITY ASSESSMENT

Figure 1 shows the workflow for the data-driven methodology to measure the true quality of a mobility-related information system. The core idea is to use vehicle fleet data to identify spatially and temporally important areas as the basis for prioritization-based subsampling strategies (PSS). This is used for the reduction in ground truth collection strategies and subsequently, quality assessment. It allows to smartly reduce ground truth collection while not missing out important areas to customer in evaluating the quality of a system.

2.1 | Acquire and process vehicle fleet data

First step was to acquire vehicle fleet data as the main source for determining the fleet data spatio-temporal density spread (see

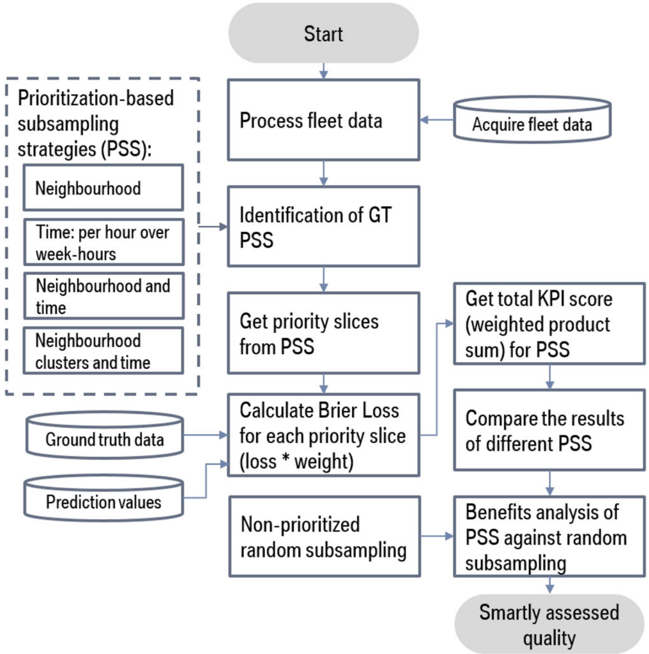


FIGURE 1 Methodology workflow

Section 3.1.1) within a city. The data was processed for geographical analysis using the geographic coordinates and timestamps. More specific processing aspects are mentioned in the strategies defined in the rest of this section.

2.2 | Identification of spatio-temporally important areas for prioritization-based subsampling strategies (PSSs)

The processed fleet data was used for identifying spatio-temporally important areas. Importance is defined by the percent volume weight (or density) of fleet data that occur within a certain area and at a specific period, hereafter referred to as ***slices***. Prioritization-based subsampling strategies (PSSs) were identified, that have different slice proportions. Various strategies were tested to have a robust experimental design setup looking at the fleet data from several perspectives. The PSSs are further elaborated in the following sections.

2.2.1 | PSS 1: Based on neighbourhoods

The first strategy was purely based on spatial slices, referred to as neighbourhoods. This strategy only considers the density of fleet data in each neighbourhood within the city over the entire study period. The spatial method considered was based on the quadkey concept [34], which is an indexing convention and unique identifier of a standard map tile at a specific zoom level. This standardized partitioning of the world map into tiles is a standard used by Microsoft's Azure Maps. The zoom level of quadkeys varies from 0 to 24, corresponding to a tile size of 40,075,017 m x 40,075,017 m to 2.39 m x 2.39 m, respectively. The finer the tile level, the lower volume of the fleet data per tile, and thereby increasing relative error. The quadkey approach is favourable to generate reproducible and comparable results for similar researches. Each quadkey equates to a slice; the densest quadkey was then considered the most important area and this was sorted from highest to lowest.

2.2.2 | PSS 2: Based on time

The time-based strategy defines slices as 168 week-hours. An hour was the selected time interval based on heuristics as it is not too small, and not too large, while maintaining interval consistency. Half-hour slices were also experimented with, but with negligible differences in the overall scores calculated in the use case in Section 3.3, hence, omitted from further analysis. The busiest week-hour is the densest slice, and thereby the most important. Typically morning and afternoon peak hours were the ones with the highest densities and after midnight hours are the quietest.

2.2.3 | PSS 3: Based on a combination of neighbourhood and time

The third strategy combines the first two. Each neighbourhood was divided into 168 h slices. The first two PSSs were on a higher aggregated level, while this PSS created lower aggregated priority. This PSS was a generic strategy that can be used in any city use case; it divided the study area spatially based on a standard quadkey approach and the week-hour basis hourly slices. This allowed for a precise identification of important areas by pinpointing neighbourhoods that are more important at specific hours during a week. The slices were sorted according to fleet volume density. Since the division was done both across neighbourhood and time, the sequence of most important slices can be from different mixtures of neighbourhoods and hour during the week. For example, the top most could be from neighbourhood A at 13:00–14:00, while the second highest could be from neighbourhood B at 8:00–9:00. Furthermore, different quadkey zoom levels indicate varying and more precise importance weighting.

2.2.4 | PSS 4: Based on neighbourhood clusters and time

Neighbourhood clusters was generated based on fleet data behaviour within the different neighbourhoods in a city. The idea was to group together neighbourhoods that have similar behaviour and can be treated as one entity. This was done by first defining the behaviour of each neighbourhood through an aggregation method of the fleet data and then performing clustering on the behavioural pattern. The behavioural modelling and clustering concept used for this paper can be found in Section 2.5. The next step was to divide the clusters into 168 h slices as previously and then sort according to density to get the importance.

2.3 | Measure the quality for different PSS by a key performance indicator (KPI)

Once the PSSs were identified and applied on the ground truth data, the different slices for each strategy were then produced. The slices were used for subsampling of the collected ground truth data. A key performance indicator was used as the quality metric. The logic behind calculating the KPI for all the strategies was to ensure that these prioritizations were consistent at different slices and measure the real quality correctly. Random sampling has in most cases been the norm [35] to reduce any biases in sampling. This paper introduces PSS as a competing method to the traditional random sampling for true quality assessment of prediction models. Moreover, an experimental design is defined to test the strategies against thousands of random sampling trials. The experimental design setup is defined to test the chances of selecting a sample, that is, areas at a specific time span that would falsely assess the quality. A popular KPI that was used is the Brier Score, as described below:

$$KPI = \frac{1}{N}\sum_{t=1}^{N}\left(p_t - o_t\right)^2, \tag{1}$$

where p is the predicted outcome, o is the observation at instance t (0 means there was no occurrence, 1 means there was an occurrence), and N is the total number of instances.

The KPI was calculated for each slice within a strategy. A total KPI score for a strategy was calculated based on the evidence-based multi-criteria decision making method called weighted sum model (WSM) as described in Equation (2). WSM was the chosen technique for its objectivity and not being prone to score skewness.

$$KPI_{PSS} = \sum_{s=1}^{N} KPI_s \times w_s, \tag{2}$$

$$w_s = \frac{PEVolume_s}{\sum_{s=1}^{N} PEVolume_s}, \tag{3}$$

where KPI_s is the KPI of a slice, w is the importance weight assigned to a slice, s is a slice within a PSS, and $PEVolume_s$ is the parking events volume at a slice

The calculation of the KPI is dependent on two variables: estimations from different types of prediction models and the strategy from different PSS. Only two weighting techniques were applied in this paper, equally weighted for all slices, which was computed by one divided by total number of slices and importance weighted based on fleet percent volume share at each slice. This was done to see the impact of weighted KPI on the overall PSS KPI, and whether the weights play a role in shifting the penalty or incentive to the important areas. After the KPIs were calculated for all PSS, the next step was to check the true quality measurement. This was done by comparing the results against the baseline, which is randomized subsampling of ground truth.

2.4 | Benefits validation of PSS against non-prioritized randomized subsampling of ground truth

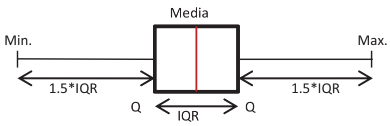
The experimental design for random subsampling of ground truth was necessary to assess and ensure the robustness of the PSS method. One objective was to ensure that if any of the PSSs are followed for ground truth collection, they can be representative of the actual quality of a prediction system. The goal of random subsampling was to generate different random slices not following fleet data density. The ideal, however, unrealistic randomized subsampling that gives the best quality measurement for a certain prediction model was also calculated as a base comparison for the benefits of the PSS implemented. This validation aimed to identify weakly designed prediction models that only perform well in rare instances. The experimental design ensured that the random trials cover the majority of the possible combinations for randomized subsampling that eventually selected all the ground truth data in different experiment setups.

The comparison of top importance-weighted fractions of the PSS with fractions of the randomized ground truth subsampling was done to compare the effects of subsample size reduction. This also provided the opportunity to check the benefits of the PSS at smaller sample sizes, which have higher relative error. It must be noted that the top importance-weights are corresponding to the fleet percent share that is attributed to a slice, and therefore not corresponding to the fraction of the ground truth observations. For instance, within the top 50th percentile importance-weighted slices, it is possible to only have a sample size of 30% of ground truth observations occurring in these specific areas and time. In summary, the following steps were followed for the benefits validation:

1. Sort the slices of each PSS based on their corresponding importance weights.
2. Take the top 30th up to 90th percentile importance-weighted slices, at 10th percentile interval steps, and calculate the KPI scores for all PSS.
3. Get the equivalent sample size % of the ground truth for the randomized selection.
4. Run n-number of trials that covers different fraction combination in consideration of the ground truth dataset size and calculate the KPI scores.
5. Get the KPI variation of the m-number of PSS.
6. Get the KPI variation of the n-number of random trials.
7. Use the interquartile range (IQR) method of outlier detection for robustness of KPI scores.

$$IQR = Q3 - Q1, \tag{4}$$

where Q3 is the third quartile value (75th percentile), and Q1 is the first (25th percentile)



$$\text{Lower bound outliers} < Q1 - 1.5 * IQR, \tag{5}$$

$$\text{Upper bound outlier} > Q3 + 1.5 * IQR. \tag{6}$$

1. Compare the score variance for random trials with the score variance for PSS.
2. Make conclusion on findings about robustness of PSS.
3. Is it feasible to safely reduce ground truth collection to only important areas and time for the quality assessment that needs to be made? Will this be representative of the true quality?

2.5 | The use case of on-street parking prediction

This paper applied the described methodology to the use case of on-street parking. The parking events dataset was used as the main source for analysis and the PSSs. The entire methodology can be applied as already described, but for PSS 4, a specific on-street parking behaviour modelling and clustering concept was used.

The neighbourhood clusters identified in this study were based on a specific parking behaviour dynamics concept taken from the study of [36] about temporal trend of parking dynamics (TTPD) inferred from parking events. TTPD is a week-hour time-series of the cumulative sum of the difference of the week-hour normalised average parked-in and parked-out events per 30-min intervals at quadkey zoom level 14. For the case of on-street parking, zoom level 14 was selected as the optimum since a more localised level would generate high relative errors given that the volume of parking events within 30-min intervals was small. Each neighbourhood at zoom level 14 has a particular normalised TTPD. These TTPDs were used as the base for clustering similar neighbourhoods. Each cluster consisted of multiple neighbourhoods and was spatially treated together, and then the cluster is divided into 168 h slices. The logic in this strategy was that, the important slices of different neighbourhoods with similar parking behaviour can be analysed on the same level and therefore combined in the cluster.

For the use case of on-street parking prediction model quality assessment, various parking prediction models were utilised to generate availability predictions. However, the model development was not of essence in this paper, and was only considered as sample models that generate adequate results to allow quality comparison between models. A number of real feature-based models and random parking prediction models were used as later described in the Section 3.3.

The code to carry out the analysis in this paper was written in Python. The main packages used were: Pandas, GeoPandas, Folium, Numpy, OSMnx, Matplotlib, Seaborn, Statsmodel, PySal, and Scikit-learn.

3 | QUALITY ASSESSMENT OF COMPETING ON-STREET PARKING PREDICTION MODELS

The application results of the methodology introduced in this paper is described in this section. The experimental design setup of the PSSs implemented is in Table 1. The experimental design was designed to cover all possible combinations of the defined spatio-temporal slices.

3.1 | Study area and description of data

3.1.1 | Study area and parking events

The study area of this paper is BMW's OSPI service area for the city of Munich, Germany. Together with the defined polygon, the on-street parking capacity of blocks or number of parking spots was also collected from BMW's parking map.

The main data source used in this study as the importance indicator was parking events (PEs). PEs data are gathered from the fleet of BMW vehicles. Hence, there is a bias towards BMW users. This is within the bound of this study since importance is relative to the OEM or the agency of concern; this means for example, the share of BMW vehicles in an area is what is defined as important for BMW, while if importance is to be defined by the city the share of BMW vehicles amongst all other vehicles need to be known to classify whether it is representative. The data collection happens at BMW's backend services which includes anonymisation according to EU defined data privacy standards. A PE is generated when a vehicle switches off or on the engine, triggering a parked-in event or parked-out event, respectively. The PE event was also post processed to contain only events within 10 meters of a street. An example of the spatial distribution of data collected can also been seen in Figure 2.

For this paper, the PE data from February 2020 to September 2020 was taken. It was observed that the PEs from Mondays to Friday evening have a similar temporal distribution with small day to day discrepancies (see Figure 3), hence, can be grouped together in later analysis [36] During a normal weekday there are peaks in the morning and afternoon, as expected since the study area is quite commercial. On weekends, the peak occurs at around noon during lunch hours and shopping before or after.

3.1.2 | Ground truth data

The ground truth (GT) data used was collected between June 2018 and October 2020. The GT dataset is used for testing the methodology. For this study, more than 20000 random observations spread across the city's service area were used in Munich.

TABLE 1 PSS experimental design setup

		Neighbourhood zoom level				TTPD cluster zoom level	Time slice
Setup #	PSS #	14	15	16	17	14	168 week-hour
1	1	x					
2			x				
3				x			
4					x		
5	3	x					x
6			x				x
7				x			x
8					x		x
9	2						x
10	4					x	x

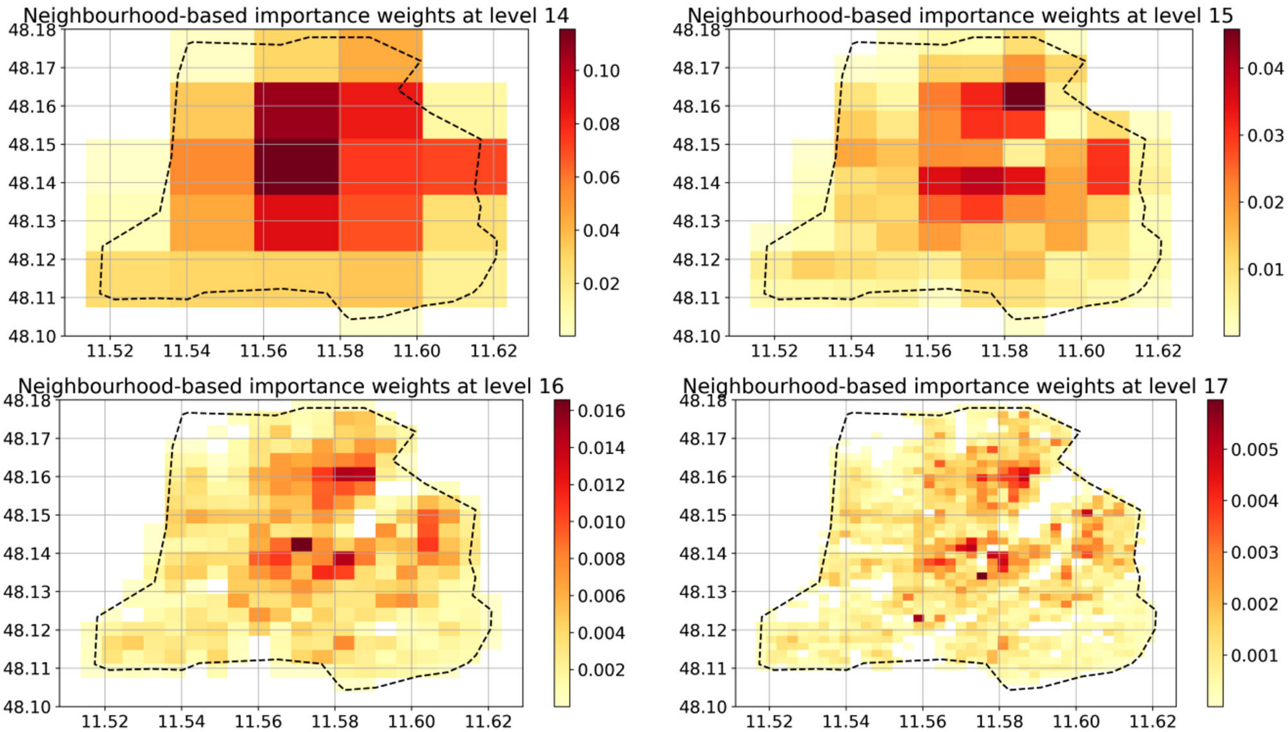


FIGURE 2 The weight importance distribution in neighbourhoods for PSS 1

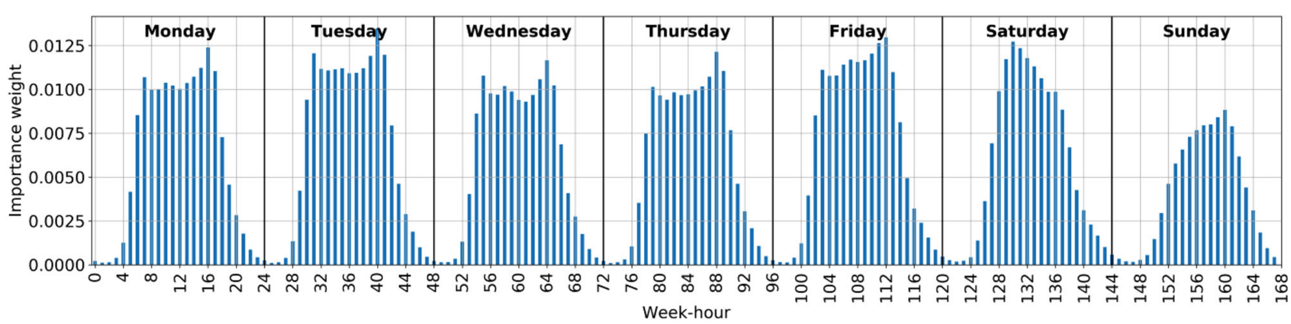


FIGURE 3 The weight distribution importance during the week-hour for PSS 2

Each observation is made on a block at the time of the test. A block is the stretch of a street measured from one intersection to the other. When at least one legal parking spot is observed on a block, this was recorded as available. Regardless of the number of open spots, for this paper, the observations were recorded as a binary outcome—available or not available.

3.2 | Spatio-temporally important areas for the use case of on-street parking in Munich

For the specific case of this paper, the volume of parked BMW vehicles is the indicator of importance. Only parking event pairs with a duration of more than 5 min were considered to eliminate noise generated by standing by cars. The hundreds of thousands of PEs that happened in Munich during the indicated collection period show the spatio-temporal importance of an area in the city. The results of the PSSs are described below.

3.2.1 | PSS 1: Based on neighbourhoods

For the neighbourhood-based prioritization (see Figure 2), the total volume of PEs in each quadkey was considered as the importance weights. Quadkey partitioning is described in Section 2.2.1. The highest and lowest quadkey zoom levels considered as a neighbourhood were level 14 (2457.6 × 2457.6 m) and 17 (250 × 250 m), respectively. These quadkey zoom levels were heuristically determined for this research as an assumption of the cruising distance range for on-street parking search. The spread of the events are mainly focused on hubs (see Figure 2) within the polygon as seen in the figures; this corresponds to the prioritized areas to focus on for the KPI calculation.

3.2.2 | PSS 2: Based on time

The global hourly based PSS applied on the PEs dataset shows (see Figure 3) that the peak importance occurs in the early mornings during the weekdays and at noon during the weekends. It is observed that on a global level, the importance by time is not that distinguishable as the weights are similar during the day hence making it difficult to prioritize. This prioritization confirms the nature of the study area as being mainly commercial and business centres. With prioritization only based on global time slices, a small trend shift of ground truth resources can be done by taking the following top prioritized hours as important: period 7:00–15:00 during weekdays, 9:00–14:00 on Saturdays, and Sundays can essentially be left out, as it is not as busy as weekdays. The observations here can change once this is looked further in detail by neighbourhood.

3.2.3 | PSS 3: Based on a combination of neighbourhood and time

PSS 3 applied to the on-street PE data provides detailed prioritized subsamples in specific areas of Munich at certain periods of time (see Figure 4). The PSS was performed for zoom levels 14 to 17, but only level 14 is discussed in this section as an example. For simplification of 14-digit labels of quadkeys in the example, a basic label encoder was used to assign a number label to each of the 23 level 14 neighbourhood quadkeys generated (see right image in Figure 4). In the final analysis of KPI scores (see Section 3.3), all levels were considered. The neighbourhoods at quadkeys 6 and 8 have the highest hourly importance contribution. It can be seen that neighbourhood 8, which is located around the central station of Munich, has the highest share, and the hourly weights are consistent throughout the day. Within the duration of 6:00 – 18:00, most neighbourhoods have stable hourly importance. In neighbourhood 14, a slight increase in importance is observed on Saturday afternoon; this neighbourhood mainly consists of shopping and dining activities. Neighbourhoods 0, 4, 10, and 18 are located at the periphery of the service area (see Figure 4) and have low volume of parking events - illustrated by light yellow indicating low importance in the upper left image in Figure 2, hence, considered as less important.

As an example, the slices that are within the top 50th percentile of importance weights are illustrated in the lower image in Figure 5. It must be noted that the weights were not normalized, and the representation in heatmap is essentially extractions from considering all slices in Figure 4. In comparison to the heatmap showing all the weights, the top 50th percentile has prioritized 539 (14.7%) slices out of 3671. And instead of looking at 23 neighbourhoods, the choices have already been reduced to 10 neighbourhoods. At higher priority areas, within top 10th percentile of importance weights, only 76 (2.0% of all) slices within 3 neighbourhoods are considered, at top 20th percentile, there are 167 (4.5%) slices in 7 neighbourhoods, at top 30th percentile, 276 (7.5%) slices within 7 neighbourhoods, and within top 40th percentile, 398 (10.8%) slices inside 7 neighbourhoods as well. Depending on the urgency to check the quality of a certain area, this PSS provides narrowed down areas and time slots that need to be checked first for quick quality measurements.

3.2.4 | PSS 4: Based on neighbourhood clusters and time

This strategy builds on the previous PSS by aggregating similar neighbourhoods. The logic behind neighbourhood clustering, as explained in [36], is to group based on same temporal trend of parking dynamics (TTPD) (see Section 2.5). The proposed method of [36] suggests using hierarchical clustering and determining the optimum number of clusters based on the silhouette score metric and the analysing its dendrogram. Applying this for the use case of on-street parking in Munich generates 7 neighbourhood clusters, where 2 (i.e. clusters 3 and 5) of them occurring at peripheries have negligible importance for BMW as they have low volume shares. Having 5 valid clusters in the study area is sufficient, as also validated in the study of [36], since the neighbourhoods within central Munich are quite similar based on the BMW PE dataset. The PSS was only applied on zoom level 14 as the considered optimal size for modelling

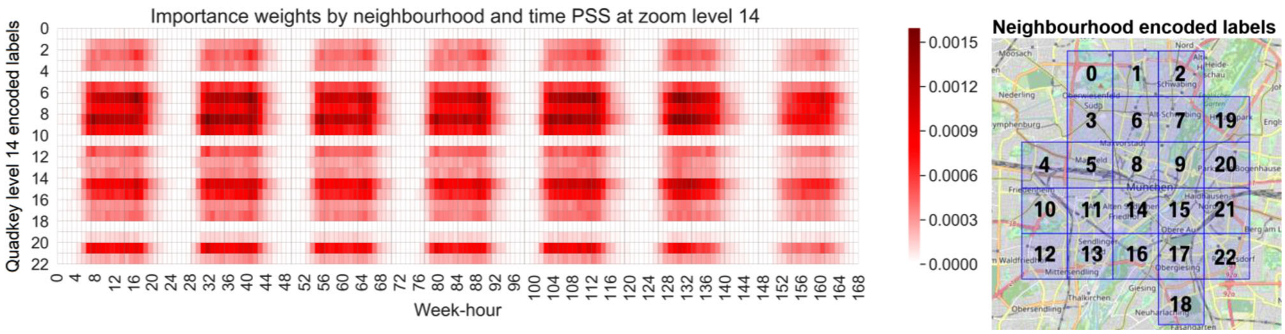


FIGURE 4 Importance weight distribution by PSS 3 on neighbourhood zoom level 14 and time (left); and encoded neighbourhood labels within Munich (right)

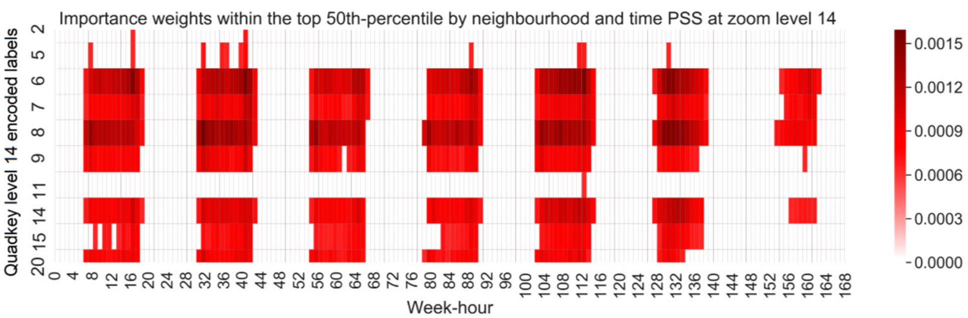


FIGURE 5 Importance weight of the same PSS 3 but for weights within top 50th-percentile

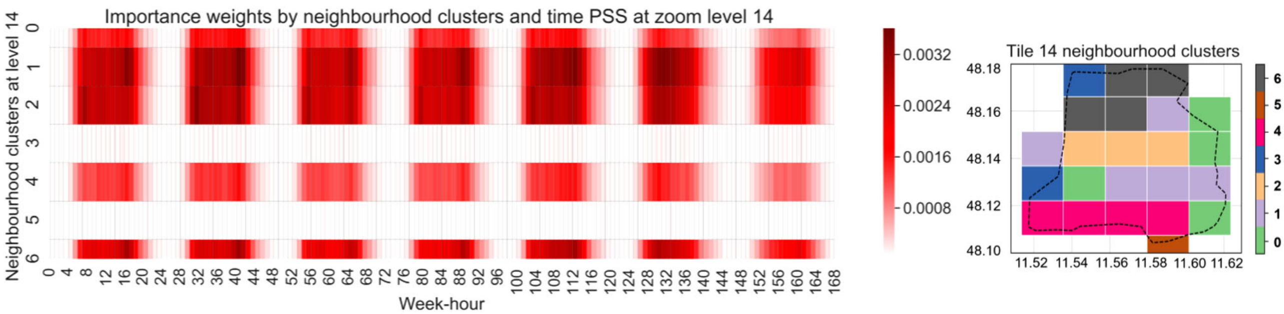


FIGURE 6 Importance weight distribution by PSS 4 on neighbourhood clusters and time (left); spatial distribution of neighbourhood clusters (right)

of temporal trends of parking dynamics (TTPD) in 15-min intervals.

The importance weights of this PSS slices are shown in Figure 6. Cluster 1 contains the majority of areas in Munich city centre and is considered important in almost all week-hours between 6:00 and 18:00, with lesser importance on Sundays. For the same period, Cluster 2 has the same stable hourly distribution with lesser magnitude in the weight. For cluster 6 the important weights are lower in the morning and intensify late afternoon and evening hours, and then fade shortly after the evening. Clusters 0 and 4 are neighbourhoods in the periphery, where the weights are lower in magnitude, but uniform during the week. The benefit of PSS 4 is that instead of being limited to certain neighbourhoods in PSS 3, similar slices can be selected from the neighbourhoods belonging to the same cluster that fits the spatio-temporal behaviour for overall ground truth strategy. The spatial distribution of the clusters showing the grouped neighbourhoods are illustrated on the right of Figure 6.

3.3 | Quality measurement of sample parking prediction models using PSSs

The generated spatio-temporally important slices from the prioritization-based subsampling strategies in Section 3.2 are now used as the input for quality measurement (see Section 2.3) of the different sample on-street parking prediction models. The Brier Loss Score was used as the KPI. This study is focused mainly on assessing the quality of various prediction models and not model improvement or development. Hence, the details of the models are not highlighted here. Only the output of the models is presented here and are evaluated

TABLE 2 Sample model algorithms

Model	Algorithm	Features	KPI	Avg. KPI Eq.	Avg. KPI Imp.
1	Xgboost	T	0.249	0.249	0.249
2	Random Forest		0.303	0.306	0.307
3	Xgboost	T, C, L	0.227	0.224	0.229
4	Random Forest		0.236	0.233	0.238
5	Xgboost	Model 3 features + hTTPD	0.228	0.226	0.231
6	Random Forest		0.231	0.231	0.235
7	Xgboost	T, h-TTPD, rt-TTPD	0.233	0.232	0.232
8	Random Forest		0.248	0.247	0.248
9	Random	Rand {0:1}	0.332	0.334	0.335
10	Optimistic Random	Rand {0.7:1}	0.273	0.267	0.273
11	Pessimistic Random	Rand {0:0.3}	0.486	0.493	0.487
12	Single Optimum Value	Average available spots	0.226	0.224	0.227

T: temporal features; C: on-street parking capacity per street; L: GT GPS location; h-TTPD: historic TTPD; rt-TTPD: real-time TTPD; Rand: random uniform between {lower limit: upper limit}; average available: average availability value of all ground truth observations for both train and test sets.

using the introduced quality assessment for comparison of the models. Twelve models were used as samples for testing the quality assessment methodology introduced in this paper. The algorithms implemented in the sample models, and some general information about the models are displayed in Table 2.

The model features and algorithms were developed with the knowledge gained from existing literature in model development for parking [7, 8, 10, 12–16]. Each model developed was either based on XGBoost [37], Random Forest [38], or random generation of probabilities. The default hyper-parameters of the model algorithms were taken without tuning. The train and test split was taken as 0.7 and 0.3, respectively, and also depending on the features that were employed. The following features in different combination were used: temporal features including time of day, month, type of day, on-street parking capacity of blocks, GPS coordinates of the ground truth observation, and temporal trend of parking dynamics (TTPD) [36].

For the calculation of the KPIs as shown in Equation (2), two weighting techniques were applied: equally weighted slices and importance weighted, respectively. Table 2 and Figure 7 display the normal KPI score without any PSS setup for each model using Equation (1), as well as the average equally and importance weighted KPI scores from the 10 PSS experimental design setups (see Table 1) using Equation (2). Models 1 to 8 use actual on-street parking related features, while 9 to 11 are random models. Model 12 is essentially an unrealistic random guesser model that only has a single optimum prediction value determined based on the expected parking availability from the ground truth data; meaning it is not forecasting, but based on all the ground truth availability, what was the average probability of finding one spot open. Nonetheless, model 12 is used as a baseline reference for comparison of quality and to test whether the quality assessment method can detect its weakness. The best models were: 3, 5, 7, and 12, whereas the worst model by large was model 11.

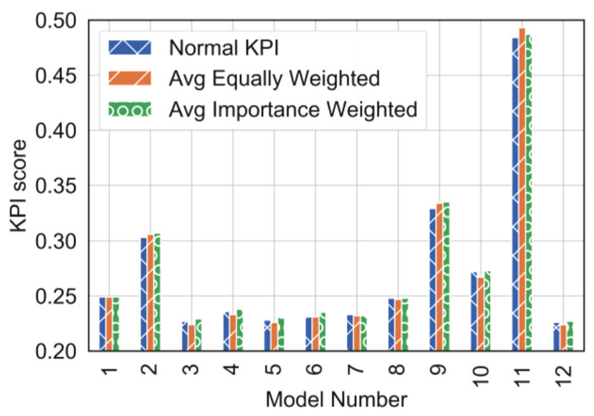


FIGURE 7 The KPI scores of each sample model

The KPI scores were calculated considering the PSSs and subsequent weightings. The range of scores per PSS can be observed in Figure 8. The figure shows the heatmaps of equally and importance weighted scores for all models against each PSS. The average scores from the heatmaps are illustrated for comparison to the normal KPI calculation in Figure 7. All feature-based models have on average a slightly worse importance weighted KPI (Brier Loss) compared to the equally weighted and normal KPI.

Figure 9 presents the average relative differences depending on the model (upper graph) and PSS (lower graph) scores, respectively. It is observed again that, on average the importance weights do not shift the scores by much from the equally weighted scores, although for each model and PSS combination the difference varies (see Figure 8). The KPI scores are on average -1.06% worse considering importance weighted for all models, while -1.07% for the PSSs. The neighbourhood-based PSSs (PSS 1) setups 1 to 4 had the largest negative relative difference between the equally and importance weighted. Setups 5

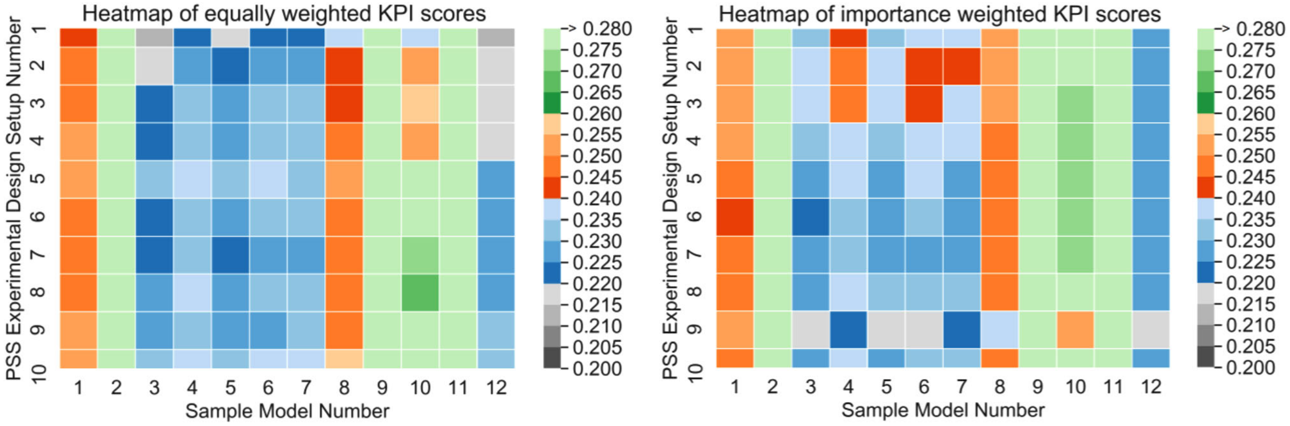


FIGURE 8 Heatmaps of equally (upper) and importance (lower) weighted KPI scores for all models considering each PSS

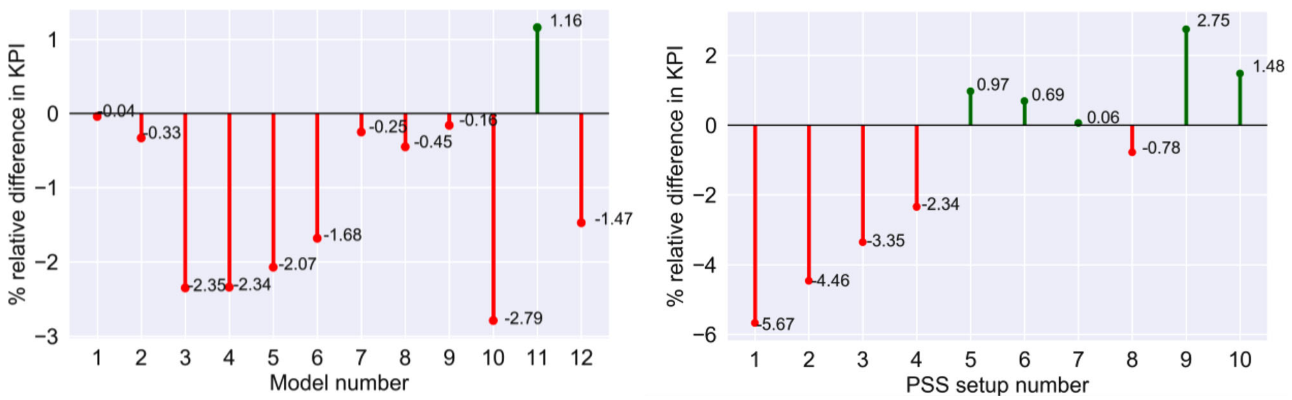


FIGURE 9 Average relative difference between equally and importance weighted KPI scores based on models (upper) or PSS setup (lower)

to 8, representing neighbourhood and time-based PSSs (PSS 3), starting at zoom level 14 and 168 week-hours incurred a positive difference but as the zoom level increased (i.e. smaller spatial scope), there was a gradual decrease in scores. For time-only-based PSS 2 setup 9, and PSS 4 neighbourhood clusters and time PSS setup 10, the calculated importance weighted scores were higher compared to the equally weighted score.

In instances when temporal aspects are considered in the experiments, the sample models' scores indicate a better performance score compared to the measured normal KPI as shown by PSS 2 setup number 9 (see Figure 9), while when a design setup includes spatial importance, the models' KPI scores are punished as shown by the average scores of PSS 1 in Figure 9 and Table 4. In the case of setup number 9, time-only-based importance assignment dilutes the goal of a prioritization-based quality assessment by disregarding the spatial factor, that is, location of parking. Hence, experimental design setup 9 is not the deciding test setup. The same can be said for PSS 4 or setup 10, in which different neighbourhoods were clustered and undermined the spatial importance of on-street parking location. PSS 4 could possibly work in a polycentric city use case, where a city has multiple equally busy centres and neighbourhoods could be more similar. However, this is not the case for Munich, as it only has one centre.

In the experimental design setups, the temporal and spatial aspects of the PSSs create a push and pull effect in the KPI measurements, thus, the average difference between equally weighted and importance weighted cannot be clearly distinguished for PSS 3 that contains setups 5 to 8 (see Figure 9), where both neighbourhood levels and the time component are considered. Further investigation shows that the reason this happens with PSS 3 is there are so many slices that are removed in the calculation of the KPI because of the lack of available ground truth for those slices (see Table 3). Table 3 displays the diminishing prioritization problem as more slices are excluded due to the lack of ground truth observations for the period of study. The lower the zoom level, the more slices are generated—this lessens the influence of prioritization-based quality assessment, unless there would be available ground truth observations at every short segment of a street within 1 h intervals. Therefore, given different PSSs, it is necessary to select a PSS that covers sufficient amount of slices generated that ensures a logical spatial aggregation and weight assignment that better represents an importance weighted or prioritization-based assessment.

TABLE 3 Excluded important slices without GT observation

Setup #	PSS #	# of Slices	# of Slices Used	Cumulative percentage of excluded important slices
1	1	23	22	0%
2		83	76	1%
3		285	232	4%
4		952	662	15%
5	3	3671	725	66%
6		12246	1264	83%
7		36756	1980	92%
8		103545	2781	97%
9	2	168	114	11%
10	4	1090	416	36%

Since the ground truth set in this study was based on random observations made throughout Munich and does not cover as many generated important slices as desired, it can only partially differentiate between equally and importance weighted KPI for PSS 3. Nonetheless, the differentiation is clear for experimental design setups 1 to 4 for PSS 1 using only spatial slices. Thus, since the results of PSS 2 and PSS 4 show the undermining of location importance, PSS 1 stands out amongst the four prioritization-based subsampling strategies tested. Furthermore, as the disparity between equally and importance weighted has been proven with PSS 1 when a significant portion of the slices are covered, the importance weighted approach is used in the benefits assessment as the basis.

Having calculated the KPI scores considering the different PSSs and weighting techniques, the next step is to check the true quality measurement. This is done by proving that this quality assessment methodology using PSSs which provides priority slices can give better insights about on-street parking prediction models as compared to doing random ground truth slices. This section covered the KPI scoring when the entire ground truth was used for the KPI measurement, while the next section covers the impact of smartly reducing ground truth data on the KPI scores.

3.4 | Benefits assessment based on comparison against non-prioritized randomized subsampling of slices

The benefits assessment (see Section 2.4) of the methodology was done by comparing the top important PSS KPI scores against the scores determined by the baseline case of non-prioritized randomized subsampling (NPRS) of ground truth. The NPRS selection was done on the slices generated from the PSS, but the importance weight was not considered, hence non-prioritized. Specifically, this section presents the impacts of top importance-based subsample reduction of ground truth size on the PSS KPI scoring and the robustness check that the methodology can eliminate the weakness of unfortunate random ground truth sampling. The ground truth sample size reduction was implemented by sorting the importance weights of the PSS slices and then taking a certain top fraction percentile. For example, using the prioritization-based reduction of GT considering only important slices of PSS setup 6 within the top 90th percentile, the GT sample size is reduced to 3563 (30% decrease) out of 5152 observations. However, if reduction was to be done randomly, 90% of the GT observations are 4637 observations. There are two reasons for the large reduction: (1) slices are only generated in areas and time frames that have recorded a parking event, hence, the GT outside of these slices are automatically disregarded as less important, in the case of the example, only 4838 observations (6% decrease) exist for PSS setup 6; and (2) there is a disproportionate distribution of the GT observations throughout the city since they were conducted randomly, and based on the performed reduction, a

TABLE 4 Percent (%) difference between weighted KPI scores and each model's normal KPI score

		Model Number												
PSS	Setup	1	2	3	4	5	6	7	8	9	10	11	12	Mean
1	1	–1.4	–4.4	–7.2	–5.8	–7.4	–8.0	–4.3	–3.3	–0.8	–15.8	11.1	–9.0	–4.7
	2	–3.0	–7.4	**–10.4**	–10.2	**–11.3**	–12.3	–8.8	–8.4	–3.1	–6.2	5.6	–4.2	–6.6
	3	–2.1	–5.6	–11.1	–8.4	–10.6	–10.8	–6.1	–6.2	–2.9	–7.6	5.8	–5.3	–5.9
	4	–3.2	–8.1	–5.3	–1.0	–4.6	–3.4	–2.5	1.7	–3.0	–6.0	4.3	–4.3	–3.0
3	5	1.1	2.9	–2.0	–3.5	–2.5	–4.7	–0.1	–1.7	3.9	–6.0	2.8	–3.3	–1.1
	6	3.0	–1.1	**1.8**	–1.2	**0.3**	–0.6	1.8	–0.4	5.7	–3.0	1.1	–1.9	0.5
	7	1.4	1.0	–1.1	–0.4	–1.4	0.2	3.5	1.1	6.4	–4.4	0.2	–2.6	0.3
	8	–0.5	–3.0	–3.3	–2.1	–4.8	–3.7	–0.7	–1.3	–0.9	–7.9	4.8	–5.1	–2.4
2	9	–3.8	3.2	**7.7**	11.3	**11.6**	11.9	7.2	5.1	1.7	12.3	–5.3	**5.9**	5.7
4	10	1.0	–0.1	**0.4**	0.7	**1.0**	–0.5	2.1	1.9	2.7	–4.8	2.6	–2.5	0.4
	Mean	–0.8	–2.3	–3.1	–2.1	–3.0	–3.2	–0.8	–1.1	1.0	–4.9	3.3	–3.2	

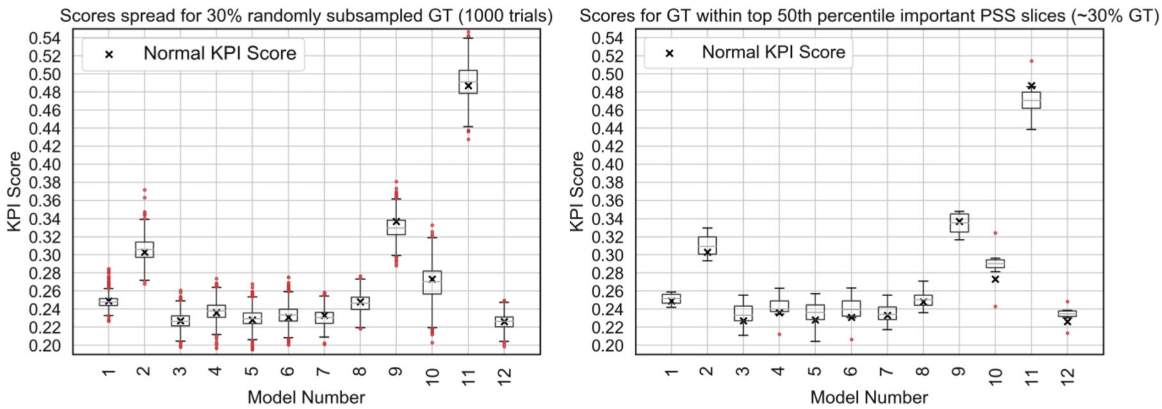


FIGURE 10 Comparison of KPI scores from NPRS against top 50th percentile important PSS

substantial amount of the collected GT were outside important areas and have sparse coverage in comparison to the number of slices for setup 6.

Further prioritization-based reduction was performed at percentile fractions ranging from 30th to 90th at 10 percentile intervals as a preliminary heuristic step. It was decided as a use case, for the main analysis here, the tiles within 50th percentile top fraction is considered. The same experimental design was setup for the NPRS. For the NPRS, at each fraction, 1000 random subsampling sets were created from the combination of 10 PSS setups and 100 unique random sampling trials. For both cases, this was done to understand the difference in the information retained about quality as compared to calculating the KPI score for the entire GT dataset. As a counterpart to the average sample size of the top 50th percentile importance fraction based on the different PSSs, only 30% GT fraction was used for NPRS. Top 50th percentile importance was selected, as the variances of scores from this fraction size onwards to 90% are relatively small.

The robustness indicator used in this analysis is the IQR method of outlier detection (see Section 2.4). This was used to measure the spread of the KPI variation for each sample model and to identify scores that were far from the central tendency. Scores that are considered as outliers are interpreted as subsampling strategies that have made an unfortunate selection of subsampling; these are not wrong, but are an indication that a strongly biased quality assessment is present. Outliers are not to be considered as part of the decisive factors. Furthermore, it can be observed on the right graph in Figure 10, the KPI scores on average are measured worse in the case of PSS compared to NPRS on the left. In the case of NPRS, 60% of scores across the first 8 feature-based models were worse than the normal KPI, while this was 69% for the PSS importance approach. This is also visible in right graph on Figure 10 as the normal KPI is consistently below the median. This signifies that, the areas and time frames belonging in the top 50th percentile important slices are harder to predict, thus, the scores are worse. This proves the need to highlight PSS important spatio-temporal slices during ground truth to measure the true quality and value of an on-street parking prediction system. Moreover, it is observed that for the pessimistic model (number 11), the scores improve in a PSS-based quality assessment (see Table 4) since the important areas are busy areas, suggesting some pessimism is necessary for a model to perform well in such areas. This is the opposite for the optimistic model number 10.

The benefits assessment proves to detect weakly designed ground truth collection strategies that give a false perception of the true quality and performance of models because of unfortunate quality testing subsampling selection. The introduced approach reveals the true performance scores. Moreover, the method can also be used to conduct a marginal benefits comparison between several competing models. This is demonstrated by investigating feature-based models 3, 5, and the top baseline unrealistic retrospective average parking availability model 12. Model 2 essentially is always just a single optimal prediction value that is equal to the average of all ground truth observations.

Models 3 and 5 both have a similar KPI value with model 12 (see Table 2) showing that these two models are high performing. To understand whether the two models are on average better than model 12, the scores are first all adjusted by applying all PSSs introduced. Model 12 had a normal KPI score ranking among the best (see Table 2) and when the scores were calculated using NPRS, the model was assessed as even better than the normal KPI in 49% of cases. Upon the selection of ground truth within the top 50 percentile important PSS slices, this occurred only in 10% of the PSS design setups (1 out of the 10) as highlighted in bold inside Table 4, and indeed detecting the model as initially falsely assessed. For models 3 and 5, the scores were better in the random NPRS scenario 52.4% and 39.2%, respectively, while as illustrated in Table 4, these feature-based models are performing better in 30% of the scores (3 out of 10 for both models) compared to the normal KPI. Based on the adjusted performances, thus, it can be gauged that the feature-based models marginally outperform the top baseline unrealistic model 12 based on a simple tally of whether the models' scores improve or get worse. In a real-world comparison, model 12 cannot exist. Hence, if the comparison is now focused on choosing between the two models 3 and 5, the next step is to select a PSS strategy that is best suited to the use case considering the available ground truth observations. As concluded in Section 3.3, PSS 1 can be the deciding factor for this study. In this case, as shown in Table 4 after adjusting the scores to setup number 2 a prioritization-based subsampling strategy (PSS 2) that focuses on neighbourhood level 15, it is concluded that model 3 is better than model 5, with KPI scores of 0.253 and 0.257, respectively, which roughly puts model 5 as 1.58% worse than model 3 after the adjusted scores.

3.5 | Synopsis of analysis

An elaborate discussion in Section 3.3 proved that PSS 1 is the most suitable for the use case presented in this paper among the 4 PSSs introduced. PSS 2 and 4 undermine spatial importance, and this is a big weakness that cannot be overcome in these PSSs. PSS 3 despite its promising approach, could not be utilised for further benefits analysis, as there was a big gap between the available ground truth observations and the number of slices generated. As presented in Table 3, although important slices can be generated using the parking events data, there is a lack of observations in order to consider the importance weights in the final score calculation for true quality. Also, it was difficult to distinguish the difference between equally and importance weighted KPI scores. However, PSS 1 does not suffer from any of these gaps, as not too many but sufficient slices are generated, that were capable of aggregating the importance and assigning reasonable weights that primarily consider the neighbourhood importance. Specifically, the most critical design setup among the 10, is setup number 2, which adjusted the models' scores on average by −6.6% as shown in Table 4. In Section 3.4, the benefits were shown by the comparison of non-prioritized randomized subsampling (NPRS) versus the PSSs. The adjustments for the worse in KPI scores were apparent and it proved that there is a need to calculate the true scores and assess models' true quality. In summary, with the application of the introduced method, it was possible to assess the true quality by reducing the ground truth subsample to areas most important to the customers, and also help decide between competing models.

4 | CONCLUSIONS AND RECOMMENDATIONS

The proposed data-driven methodology in this paper has shown that it is possible to smartly reduce ground truth and still assess the true quality of different prediction models by multiple prioritization-based subsampling strategies (PSSs). The approach automatically identifies important neighbourhoods (space) and time periods, called slices, based on the volume share of the fleet's parking events within them. Different PSSs were introduced that can be applied to any type of fleet data prioritization strategy. For the use case of on-street parking information (OSPI), the method was applied using the parking events dataset of Munich, Germany.

The methodology benefits assessment confirms that, the prioritization-based technique is capable of identifying false assessment of models. This was evaluated based on a comparison with non-prioritized randomized subsampling (NPRS) on a 30% fraction of the ground truth dataset. The NPRS approach was done to quantify the chances of misfortunately randomly selecting areas and time periods that do not necessarily represent the true quality. This was accomplished by assessing the quality metric scores at the automatically defined slices across the 10 PSS design setups that were tested. The PSS approach considered the top 50% important slices as the subsample to assess the true quality of the different OSPI models. In majority of the cases, the measured scores at important slices that are more valuable to potential customers, the models performed worse in comparison to NPRS. This implies that assessing the quality at the defined important slices must be checked first before other areas and time periods are observed. The prioritization method then immediately gives a robust first impression of a model's performance.

In conclusion, it is possible to make mistakes of wrongly assessing the true quality of a model when the ground truth data is collected randomly. The usage of the prioritization-based quality assessment is that, collectively, the PSSs can robustly evaluate the performance of a mobility-related prediction model, where it matters most to the users of the system. The methodology also allows the quality managers to gain first valuable insights fast at a lower cost with less ground truth needed. Thus, the introduced methodology in this study can directly be used by companies that are maximizing their resources for quality testing of mobility-related information systems.

The next possible directions of this research are to conduct a comprehensive study on the optimized minimum fraction of ground truth required for the true quality assessment check, the application of the methodology on other mobility use cases, and the extension of prioritization-based subsampling strategies using other factors such as the density of points-of-interest (POIs) or local contextualized information and so on. The prediction models presented in this paper were only used as examples to demonstrate the capability of the quality assessment methodology introduced in this research. As research continuous, there are plans to do a study on model development and improvement.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the support of INRIX in gathering ground truth data, which was a fundamental asset used in this research.

CONFLICT OF INTEREST

There are no conflict of interest issues related to any author.

DATA AVAILABILITY STATEMENT

The data is only accessible per request at BMW, which requires approval and authorization from the persons in charge.

ORCID

Syrus Gomari https://orcid.org/0000-0001-5179-2245
Constantinos Antoniou https://orcid.org/0000-0003-0203-9542

REFERENCES

1. Bogenberger, K., Weikl, S.: Quality management methods for real-time traffic information. Procedia-Social Behav. Sci. 54, 936–945 (2012).
2. He, Y., Csiszar, C.: Quality assessment method for mobility-as-a-service based on autonomous vehicles. In: Proceedings of International Conference on Transportation and Traffic Engineering(ICTTE), pp. 901–910. ACM, New York (2018)
3. Mugion, R.G., Toni, M., Raharjo, H., Di Pietro, L., Sebathu, S.P.: Does the service quality of urban public transport enhance sustainable mobility? J. Clean. Prod. 174, 1566–1587 (2018).
4. Guimaraes, T., Armstrong, C.P., Jones, B.M.: Evolving a comprehensive measures for system quality. Decision Sciences Department, Tennessee Technological University (2008)
5. Friedrich, T., Krejca, M.S., Rothenberger, R., Arndt, T., Hafner, D., Kellermeier, T., Krogmann, S., Razmjou, A.: Routing for on-street parking search using probabilistic data. AI Commun. 32(2), 113–124 (2019).
6. Hampshire, R.C., Shoup, D.: What share of traffic is cruising for parking? J. Transp. Econ. Policy 52(Part 3), 184–201 (2018)
7. Gkolias, K., Vlahogianni, E.I.: Convolutional neural networks for on-street parking space detection in urban networks. IEEE Trans. Intell. Transp. Syst. 20(12), 4318–4327 (2019).
8. Jomaa, H.S., Grabocka, J., Schmidt-thieme, L.: A hybrid convolutional approach for parking availability prediction. In: Proceedings of 2019 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. IEEE, Piscataway (2019)
9. Pel, A.J., Chaniotakis, E.: Stochastic user equilibrium traffic assignment with equilibrated parking search routes. Transp. Res. Part B Methodol. 101, 123–139 (2017).
10. Rong, Y., Xu, Z., Yan, R., Ma, X.: Du-parking: Spatio-Temporal big data tells you realtime parking availability. In: Proceedings of KDD 18: The 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 646–654. ACM, New York (2019),
11. Caicedo, F., Robuste, F., Lopez-Pita, A.: Parking Management and modeling of car park patron behavior in underground facilities. Transp. Res. Rec. J. Transp. Res. Board 1956(1), 60–67 (2006).
12. Pflügler, C., Köhn, T., Schreieck, M., Wiesche, M., Krcmar, H.: Predicting the availability of parking spaces with publicly available data. In: Proceedings of INFORMATIK 2016, pp. 361–374. Gesellschaft für Informatik, Bonn (2016)
13. Dosch, C., Farghal, M., Kapsecker, M., Lorenz, C., Mosharafa, A.: Parking Prediction in the City of Melbourne (2017)
14. Shao, W., Zhang, Y., Guo, B., Qin, K., Chan, J., Flora, D.: Parking availability prediction with long short term memory model. In: Proceedings of International Conference Green, Pervasive, and Cloud Computing(GPC 2018), pp. 124–137. Springer, Cham (2018)
15. Yang, S., Ma, W., Pi, X., Qian, S.: A deep learning approach to real-time parking occupancy prediction in spatio-temporal networks incorporating multiple spatio-temporal data sources. Transp. Res. Part C 107, 248–265 (2019).
16. Ji, Y., Tang, D., Blythe, P.T., Guo, W., Wang, W.: Short-term forecasting of available parking space using wavelet neural network model. IET Intell. Transp. Syst. 9(2), 202–209 (2015).
17. Xiao, X., Jin, Z., Hui, Y., Xu, Y., Shao, W.: Hybrid spatial–temporal graph convolutional networks for on-street parking availability prediction. Remote Sens. 13(16), 1–20 (2021).
18. Balmer, M., Weibel, R., Huang, H.: Value of incorporating geospatial information into the prediction of on-street parking occupancy–A case study. Geo-Spatial Inf. Sci. 24(3), 438–457 (2021).
19. Awan, F.M., Saleem, Y., Minerva, R., Crespi, N.: A comparative analysis of machine/deep learning models for parking space availability prediction. Sensors 20(1), 322 (2020).
20. Monteiro, F.V., Ioannou, P.: On-street parking prediction using real-time data. In: Proceedings of 2018 IEEE 21st International Conference on Intelligent Transportation Systems (ITSC '18), pp. 2478–2483. IEEE, Piscataway (2018)
21. Yang, S., Qian, Z.: (Sean): Turning meter transactions data into occupancy and payment behavioral information for on-street parking. Transp. Res. Part C Emerg. Technol. 78, 165–182 (2017).
22. Stenneth, L., Wolfson, O., Xu, B., Yu, P.S.: PhonePark: Street parking using mobile phones. Proceedings of the 2012 IEEE 13th International Conference on Mobile Data Management, MDM, pp. 278–279. IEEE, Piscataway (2012)
23. Xu, B., Wolfson, O., Yang, J., Stenneth, L., Yu, P.S., Nelson, P.C.: Real-time street parking availability estimation. In: Proceedings of the 2013 IEEE 14th International Conference on Mobile Data Management, pp. 16–25. IEEE, Piscataway (2013)
24. Ye, X., Wang, J., Wang, T., Yan, X., Ye, Q., Chen, J.: Short-term prediction of available parking space based on machine learning approaches. IEEE Access 8, 174530–174541 (2020).
25. Shi, F., Wu, D., Liu, Q., Han, Q., Mccann, J.A.: ParkCrowd : Reliable crowdsensing for aggregation and dissemination of parking space information. IEEE Trans. Intell. Transp. Syst. 20(11), 4032–4044 (2018).
26. Di Martino, S., Origlia, A.: Exploiting recurring patterns to improve scalability of parking availability prediction systems. Electron. 9(5), 1–19 (2020).
27. Bibi, N., Majid, M.N., Dawood, H.: Automatic Parking Space Detection System. In: Proceedings of 2017 2nd International Conference on Multimedia and Image Processing (ICMIP 2017) - IEEE, pp. 11–15. IEEE, Piscataway (2017)
28. Bock, F., Di Martino, S.: How many probe vehicles do we need to collect on-street parking information? In: Proceedings of 5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems MT-ITS 2017, pp. 538–543. IEEE, Piscataway (2017)
29. Coric, V., Gruteser, M.: Crowdsensing maps of on-street parking spaces. In: Proceedings of 2013 IEEE International Conference on Distributed Computing in Sensor Systems (DCOSS), pp. 115–122. IEEE, Piscataway (2013)
30. Arab, M., Nadeem, T.: MagnoPark: Locating on-street parking spaces using magnetometer-based pedestrians' smartphones. In: Proceedings of 2017 14th Annual IEEE International Conference on Sensing, Communication, and Networking, SECON 2017, pp. 1–9. IEEE, Piscataway (2017)
31. Mantouka, E.G., Fafoutellis, P., Vlahogianni, E.I.: Deep survival analysis of searching for on-street parking in urban areas. Transp. Res. Part C Emerg. Technol. 128, p. 103173 (2021).
32. Ajeng, C., Gim, T.H.T.: Analyzing on-street parking duration and demand in a Metropolitan City of a developing country: A case study of Yogyakarta City, Indonesia. Sustainability 10(3), 591 (2018).
33. Navarro-B, J.-E., Gebert, M., Bielig, R.: On automatic extraction of on-street parking spaces using park-out events data. In: Proceedings of 2021 IEEE International Conference on Omni-Layer Intelligent Systems (COINS), pp. 1–7. IEEE, Piscataway (2021)
34. Microsoft: Zoom levels and tile grid. https://docs.microsoft.com/en-us/azure/azure-maps/zoom-levels-and-tile-grid?tabs=csharp#quadkey-indices%5C. Accessed June 2019
35. Singh, S.: Simple random sampling. In: Advanced Sampling Theory with Applications, pp. 71–136. Springer, Dordrecht (2003)
36. Gomari, S., Knoth, C., Antoniou, C.: Cluster analysis of parking behaviour: A case study in Munich. Transp. Res. Procedia 52(2021), 485–492 (2021).
37. Chen, T., Guestrin, C.: XGBoost : A scalable tree boosting system. In: Proceedings of KDD '16: 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794. ACM, New York (2016)
38. Fawagreh, K., Gaber, M.M., Elyan, E.: Random forests: From early developments to recent advancements. Syst. Sci. Control Eng. An Open Access J. 2(1), 602–609 (2014)

How to cite this article: Gomari, S., Knoth, C., Antoniou, C.: Prioritization-based subsampling quality assessment methodology for mobility-related information systems. IET Intell. Transp. Syst. 16, 602–615 (2022). https://doi.org/10.1049/itr2.12160

C Gomari et al. (2023). Development of a data-driven on-street parking information system using enhanced parking features

Reference: S. Gomari, R. Domakuntla, C. Knoth, and C. Antoniou, "Development of a data-driven on-street parking information system using enhanced parking features". *IEEE Open J. Intell. Transp. Syst.,* vol. 4, pp. 30-47, 2023. https://doi.org/10.1109/OJITS.2023.3235898

The summary of this study can be found in Section 4.

Credit contribution statement:
The paper contributions per author are listed in bullets below:

- Syrus Gomari: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing-original draft, Writing-review & editing
- Rohith Domakuntla: Data curation of parking behaviour change detection, Investigation of parking behaviour change detection
- Christoph Knoth: Funding acquisition, Methodology, Project administration, Resources, Supervision, Writing-review & editing
- Constantinos Antoniou: Formal analysis, Methodology, Supervision, Writing-review

Received 29 November 2021; revised 21 May 2022 and 19 October 2022; accepted 30 December 2022. Date of publication 10 January 2023; date of current version 26 January 2023.

Digital Object Identifier 10.1109/OJITS.2023.3235898

Development of a Data-Driven On-Street Parking Information System Using Enhanced Parking Features

SYRUS GOMARI [1,2], ROHITH DOMAKUNTLA [1], CHRISTOPH KNOTH [3], AND CONSTANTINOS ANTONIOU [1]

[1] Chair of Transportation Systems Engineering, TUM School of Engineering and Design, Technical University of Munich, 80333 Munich, Germany

[2] Team Connected Parking, Technical Product Design Location Based Services, BMW Group, 80788 Munich, Germany

[3] Analog and RF Verification, Infineon Technologies AG, 81726 Munich, Germany

CORRESPONDING AUTHOR: S. GOMARI (e-mail: syrus.gomari@gmail.com).

ABSTRACT On-street parking information (OSPI) systems help reduce congestion in the city by lessening parking search time. However, current systems use features mainly relying on costly manual observations to maintain a high quality. In this paper, on top of traditional location-based features based on spatial, temporal and capacity attributes, vehicle parked-in and parked-out events are employed to fill the quality assurance gap. The parking events (PEs) are used to develop dynamic features to make the system adaptive to changes that impact on-street parking availability. Additionally, a parking behavior change detection (PBCD) model is developed as an OSPI supplementary component to trigger potential parking map updates. The evaluation shows that the developed OSPI availability prediction model is on par with state-of-the-art models, despite having simpler but more enhanced and adaptive features. The foundational temporal and aggregated spatial parking capacity features help the most, while the PE-based features capture variances better and enable adaptivity to disruptions. The PE-based features are advantageous as data are automatically gathered daily. For the PBCD model, impacts by construction events can be detected as validation. The methodology proves that it is possible to create a reliable OSPI system with predominantly PE-based features and aggregated parking capacity features.

INDEX TERMS Change detection, connected vehicles, geospatial analysis, intelligent transportation systems, machine learning, parking, vehicle navigation.

I. INTRODUCTION

A. BACKGROUND

Vehicles cruising for parking are estimated to contribute to 30% congestion within a transport network [1]. This causes noise, air pollution, and travel time delays. As a parking management measure, cities have invested in parking guidance signs to direct cars to primarily off-street parking lots and multi-story car parks. Comparable systems have also recently been developed for finding parking spots on the streets, denoted as on-street parking information (OSPI). One of the benefits of such services is reduction of traffic congestion caused by cruising for a parking space [2], [3], [4], [5].

Connected intelligent transport systems (C-ITS), such as OSPI, have the potential to efficiently and better distribute vehicles within a transport network as they search for parking. Reliability and quality of such information systems must be ensured to offer dependable services that contribute to helping people make better decisions on how to navigate inside the city or whether to even use a car or not.

The content of state-of-the-art OSPI systems are mostly developed using complex engineered features and machine learning techniques [2], [3], [5], [6], [7], [8], [9]. The main

The review of this article was arranged by Associate Editor Emmanouil Chaniotakis.

difference between the models available are the data gathered for training the models and the incorporated features in the models. The differences in input data play a major role in the reliability and quality. The quality of the information provided by such systems are validated by the comparison of observed on-site data against the prediction model estimates.

B. PROBLEM STATEMENT WITH THE CURRENT SYSTEMS

Continuous manual ground truth collection for parking information systems is costly. The level-of-service and reliability remains an on-going challenge within the industry. This is attributed to difficulties in gathering accurate yet scalable data with adequate spatial and temporal coverage relative to the localized information needed. Many researches have used sensor data to develop models, but as stated in [6], these incur high costs of installation. Further, maps are usually only updated every quarter [10] as it is likewise a costly process to do so. This can be problematic when there are mid- to long-term changes that last from a few weeks to permanently. This is especially true for the case of on-street parking, since searching in an area that has obstructed parking could considerably increase the parking search time. For a parking service the sooner the changes are known, the better a service can be and parking availability models can be updated as well. As such, the goal is to provide the same quality of a prediction model with a scalable set of features based on sound domain knowledge to engineer features that rely on smart systems and less on-site surveyors.

This issue is partially tackled in this study with the use of real-time and readily available parking events data, which can be used to engineer added-value features to an OSPI service. Additionally, the same dataset could be used to specifically help parking maps be adaptive with the use of parking behavior change detection trigger.

C. CONTRIBUTIONS AND MAIN OBJECTIVE

The contributions of this research are as follows:

- The value discovery in vehicle parking events as a source to extract a wide range of features to enhance an on-street parking information system. These features include variations of hourly to weekly moving averages of time-series parked-in and parked-out data. The proposed OSPI system also has a parking events-based adaptive feature with a supplementary parking behavior change detection (PBCD) feature that is more dynamic as it can detect mid- to long-term (i.e., more than 10 days) static anomalies, closures, and disruptions signaled by the drop of parking events caused by construction obstructions, rule changes, or significant infrastructural changes, among others. These detections, essentially, convert predictions to zero to indicate unavailability of parking on top of an alert trigger to drivers to flag and confirm potential changes relating to on-street parking provisions and as an alert for the evaluation of the OSPI system. To the best knowledge of the authors, currently, there are no systems in practice or in research that updates their maps and predictions using such a dataset.
- The domain knowledge of the authors enhanced engineering of parking features from the parking events data and spatial parking capacity data previously unknown. Engineered valuable features from simple spatial capacity features that are easy to collect and prepare as input for an on-street parking availability model. Simple spatial on-street parking capacity features become more valuable when aggregated on a higher neighborhood (i.e., quadkey) level. Rather than just having the capacity information on a street-level, aggregation on a higher level can capture variances that supplements the variances captured through the street-level capacity feature.
- This proposed OSPI system can replace a system which solely relies on a prediction model that depends on continuous expensive parking availability features to keep the information system up-to-date. Shifting away from such a system lessens the cost associated with manual ground truth collection and allows faster scaling.

As opposed to many researches that have been done using complex models to create parking prediction models, this study aims to use less time-intensive machine learning algorithms that are easier to comprehend, interpret, and implement. Thus, the focus is on utilizing domain knowledge to engineer features to improve an OSPI system while using a readily available machine learning algorithm that only needs to be trained and hyperparameters-tuned. Developing a new machine learning algorithm is out of scope.

The paper is organized as follows. Related literature is described in Section II. Section III covers the main discussions of this paper. The data and study area are introduced in Section III-A. The elaboration of the development methodology of the OSPI system is presented in Section III-A1. Section IV presents the supplemental OSPI feature developed with the parking behavior change detection methodology that represent the dynamicity of the proposed OSPI system. The specifics regarding the features, algorithm hyperparameters, and the evaluation of the models are described in Section V. Section VI gives concluding remarks and some recommendations.

II. RELATED LITERATURE

The proposed approach in this study focuses on developing a data-driven OSPI system focused on valuation generation from different data sources while using prominent machine learning algorithms as the different baseline models. The logic behind this is that domain knowledge in parking can enhance the model developed. The literature review in this section is subdivided to the ground truth data used for validation in parking studies, the supplementary data used to engineer features that are not dependent on ground truth data, the popular parking prediction machine learning models that have been used in research, and the usage of parking behavior change detection models in OSPI systems. The review here mainly focuses on on-street parking.

A. GROUND TRUTH DATA USED FOR VALIDATION OF PARKING STUDIES

Most state-of-the-art on-street parking availability models developed today use a diverse range of data sources. This can be classified to two: data only used for feature engineering and ground truth data primarily used for training, testing, and validating. The latter can also be used for feature engineering.

Different types of ground truth data exist for validation of on-street parking prediction models. Some have used parking sensors in researches [8], [9], [11], [12], [13], [14], [15]. Some [9], [16], [17] have also used parking meter payments or mobile payments [3], [18] as a type of sensor to infer parking availability. A study also used costly labor-based manual observations for validation [19]. Another line of research [2], [20] have used images and videos from the camera of a moving vehicle to identify on-street parking spaces by processing these through some machine learning image recognition algorithm. Some researchers also employed crowd-sensing information by equipping probe vehicles with on-board sensors, cameras, or ultrasonic sensors [20], [21]. There are also studies who have explored the usage of crowd-sourcing data from smartphones or Global Positioning System (GPS) devices [18], [20], [22], [23].

Most of the ground truth data sources abovementioned have been studied to replace the longstanding industry practice that is still primarily based on manual ground truth collection to the best knowledge of the authors. The main reason is, each alternative ground truth is either limited in scope in different cities, such as street parking sensors and meters, and/or is unscalable. If different ground truth sources are used for each model in each city, this can be problematic as it will increase development costs of a system. Hence, the dependence on reliable manual observation.

An apparent gap that exists in all studies is that they have not tested these other ground truth sources to instead support manual ground truth to reduce frequency of manual observations required in practice. That is, the training of a model can be based on the manual observations, and the coverage-limited data gathered can be used as updates to the system since it is automatically collected albeit being sparse in space and time. The focus of the studies has been to completely replace them without direct comparisons against models that rely completely on manually gathered ground truth data.

In this study, the authors propose to use the cheaply and automatically collected sparse parking events data as a source to support manual ground truth collection and reduce the frequency of collection.

B. FEATURES IN PARKING BEHAVIOR AND PARKING PREDICTION STUDIES

On-street parking behavior and prediction studies have used a variety of features for their models. Common practice is to use the data as is as a feature and do feature engineering in this data to possibly capture different variances to better predict the target value. Two common features in research are temporal and spatial features mainly taken from the ground truth parking availability data that inherently has a location and time component. This typically is the composition of a baseline model's feature set. A few studies incorporated traffic data in their parking prediction models [6], [15], [24], [25] – this can be in the form of speed or their own engineered features to get traffic congestion indices. Some studies also have used parking-specific influencing factors such as parking pricing to understand changes in parking occupancy [26], [27]. Such factors can be used on street-level features. Another study used on-street parked out events to classify legal and illegal parking spots in the city [28]. Floating car data is another indirect source to infer parking behavior [1], [29]. Weather data has been proven by many studies to either help make prediction models or understand parking behavior [6], [9], [25]. Some other features that are also incorporated include map-related features such as street length, landuse, and points-of-interest (POI) data regarding shops, parking facilities, [1], [5], [7], [15], [19]. A few studies also included special events [5], [6]. A particularly interesting approach was done using survey data by studies like that done by Google's research team, where they asked about the subjective difficulty of parking in one's search area [30].

All studies besides a few do not give details regarding the features engineered. Particularly, a gap was observed in further aggregating simple features such as street capacity. This is typically done on temporal features, where moving averages or aggregation on various intervals are incorporated, but spatial aggregation has not been explored much according based on the literature reviewed. Studies also primarily focus on developing better algorithms than focusing on the usage of domain knowledge for feature engineering to improve their parking prediction models.

C. POPULAR PARKING PREDICTION MODELS

Parking prediction modelling studies have become popular in the last years since the hype of big data. There is a wide range of machine learning models that have been employed by researchers in the last few years. The following models have been tested in the reviewed studies: clustering [15], [21], [31], different linear regression algorithm like Lasso, Ridge, or basic linear regression [32], vector spatio-temporal autoregression [13], ARIMA [25], Support Vector Machine classifier [33], decision tree [15], [28], random forest [7], Support Vector Regression [14], [25], and tree-based algorithms like Gradient Boosting Regression Tree (GBRT) [15], [34] among others. Despite longer run times and in the hopes that unsupervised learning can enhance models, many studies have utilized deep learning approaches using neural networks like multi-layer perceptron [15], [35], CNN, Hybrid CNN, Graph CNN, RNN, LSTM, [2], [3], [6], [8], [9], [25]. Another one used logistic probability distribution and aggregating over all the observations [16]. XGBoost [36], one of the currently popular algorithms in various fields that uses a type of gradient tree boosting system that resembles an ensemble tree model, was employed by several studies [3], [7], [24] that showed the most promise in the use case of our proposed system as well. Google's research team used a single layer regression and feed forward deep neural network [30] for estimating difficult of parking using mainly Google maps travel data.

D. PARKING BEHAVIOR CHANGE DETECTION MODELS

There are no known studies that specifically use parking events to determine potential changes in parking behavior associated with longer term static changes like in rules and restrictions, constructions, or infrastructural changes. There was one study by [37] that used sensor data as well for detection of unusual patterns and infer it to any possible disturbances to parking location or sensors. Reference [28] used park-out events to detect anomalies with regards to classifying legal and illegal parking spots in relation to their map.

Majority of the studies have relied on explicit usage data input from on-street parking sensors or apps, while implicit recognition of parking occupancy has not be widely used [6]. In our study, we employ user data from parked-in and parked-out events to partly infer parking availability in conjunction with other features. The aim is to combine these data with readily available machine learning algorithms that could compete on the same level as commercial OSPI models. Although we aim to provide real-time updates to the model through introducing parking events-based features, parking events cannot be used for validation as half of the picture is missing. Fully occupied streets (true negatives) and streets that were predicted to have parking but did not (false positives) also cannot be validated with parking events, hence, it was used primarily as a source to engineer features. Nonetheless, as an added component to an OSPI system, the parking events data is also utilized to provide map triggers about potential on-street parking behavior changes that are caused by long term external factors such as construction.

III. DEVELOPMENT METHODOLOGY OF A DATA-DRIVEN ON-STREET PARKING INFORMATION (OSPI) SYSTEM

A. DESCRIPTION OF DATA USED

This section describes the data that were used in this study for training and evaluation of the model. The data that were used to extract features from are also presented. BMW's OSPI service area for the city of Munich, Germany was the chosen city use case for this paper.

The data sources are only described on a high-level to not violate BMW data confidentiality policies. Absolute numbers and descriptive statistics cannot be elaborated upon. Nonetheless, details relevant for the development of an OSPI system are described here.

1) PARKING EVENTS

Feature extraction from parking events (PEs) is one of the main contributions of this paper. Parking events (PEs) data are gathered from the fleet of BMW vehicles and are collected at BMW's backend data center. Hence, there existence of the bias towards these users. All parking events adhere to anonymization according to EU defined data privacy standards. A PE is generated when a car engine switches off or on, corresponding to a parked-in event or parked-out event, respectively (see Fig. 2). The PE event was also post processed to contain only events within the proximity of a street. Further details about the nature of the parking events dataset are discussed in [38] and [39]. As opposed to studies reliant on ground [7], [8] which cover only certain parts of a city, this research aims to utilize parking events as floating sensors.

Hundreds of thousands of parking events data used was gathered between May 2019 and October 2020 with a gap between October 2019 and February 2020.

2) GROUND TRUTH OBSERVATIONS

The ground truth (GT) data used was collected between May 2019 and October 2020. The GT observations were used for training and testing the models developed. For this dataset, the sparse data collection strategy (i.e., where, when, and how much data) was beyond the control of the authors. In the validation phase and the final scoring phase, a prioritization-based quality assessment [42] is used to adjust the scores depending on the amount of parking events that occurred in each spatio-temporal cell. This helps eliminate unimportant hours. In this study, more than 10000 random walk observations were made within the central area of Munich, Germany. Each recorded observation was made on a street block (i.e., intersection to intersection) at the time of collection. When at least one legal parking spot is observed on a block, this was recorded as available. Regardless of the number of open spots, the observations were recorded as a binary outcome – available (1) or not (0). Most foundational and important features are extracted from these observations. Among others, this includes spatial and temporal features further described in Table 1. In Fig. 1, the average parking availability aggregated on quadkey level 14 over a period of 168 week-hours is illustrated. Since observations were mostly random, there is an uneven distribution of collection throughout the city. Fig. 3 represents the spatial distribution of each observation. The average parking availability in the entire study area is 0.56. Central busy areas such as neighborhoods 6, 8, 14, and 16 (see Fig. 1) are more difficult to predict compared to the periphery.

A time series split (i.e., temporally sorted) cross validation was implemented for training and testing. In this case, testing sets here are considered the evaluation sets as well. The data is split into three equal partitions to conduct two

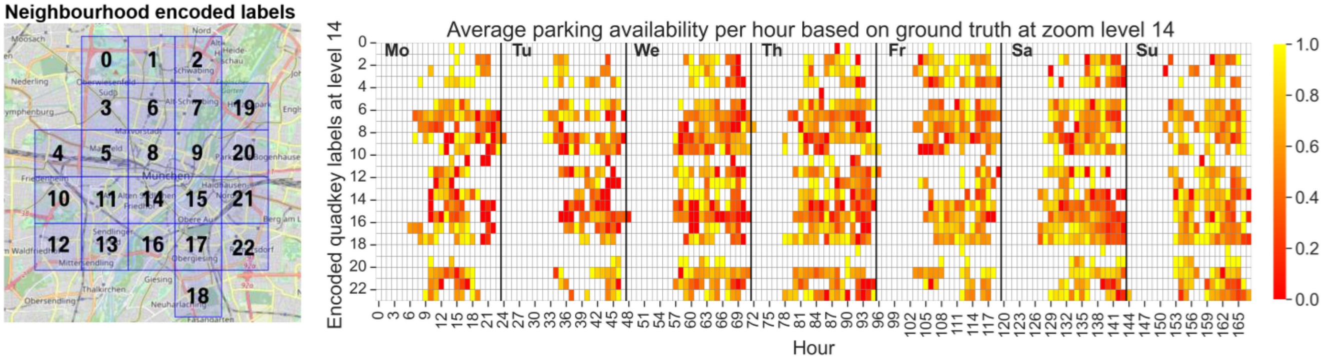


FIGURE 1. The average parking availability aggregated over 168 week-hours at zoom level 14 in Munich's study area.

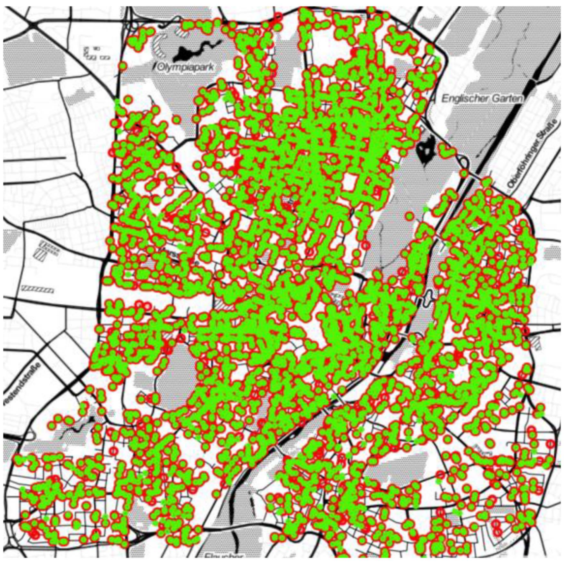


FIGURE 2. Paired parking events in Munich for one day. Green is for parked-out events and red is for parked-in events.

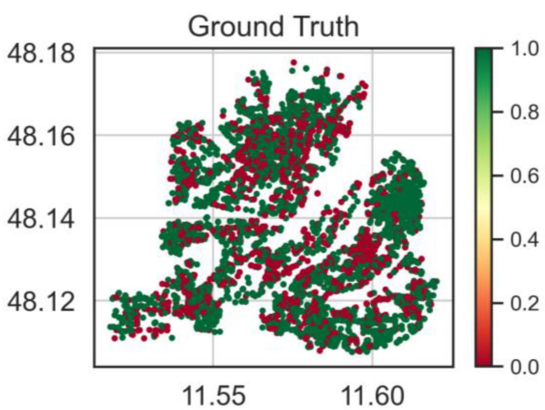


FIGURE 3. Spatial distribution of ground truth observations.

cross validation iterations. In the first iteration, the first 33% of the ground truth observations are used for training the model, and the next 33% used for evaluation. The second iteration takes the first 66% for training and the last 33% for evaluation.

FIGURE 4. Time series split cross-validation (CV) train and test sets.

3) TRANSPORT NETWORK FOR ON-STREET PARKING

BMW's transport network consists of on-street blocks as defined above. The main feature used from here is the number of legal parking spots or on-street parking capacity of each block.

4) OTHER MAP DATA AND WEATHER DATA

To further enhance the features of the model, map data regarding construction were requested from HERE maps (2021). Furthermore, open weather data were downloaded from Deutscher Wetterdienst (2021). Only temperature and rainfall data were used in the models.

B. METHODOLOGICAL FRAMEWORK FOR OSPI DEVELOPMENT

The core feature of an OSPI system is the provision of an availability prediction to show the users the chances or difficulty of finding a parking spot in certain areas at given time periods. Particularly, the availability model that was developed in this study, as part of its novel contribution, uses mainly parking events-based features, which are dynamic in nature and uphold or improve the performance of a model. Despite the unbalanced nature of the PE dataset, the goal was to develop a model that is up to the level of commercial models. The PE dataset is unbalanced as it only provides information about open spots and occupied parking spots

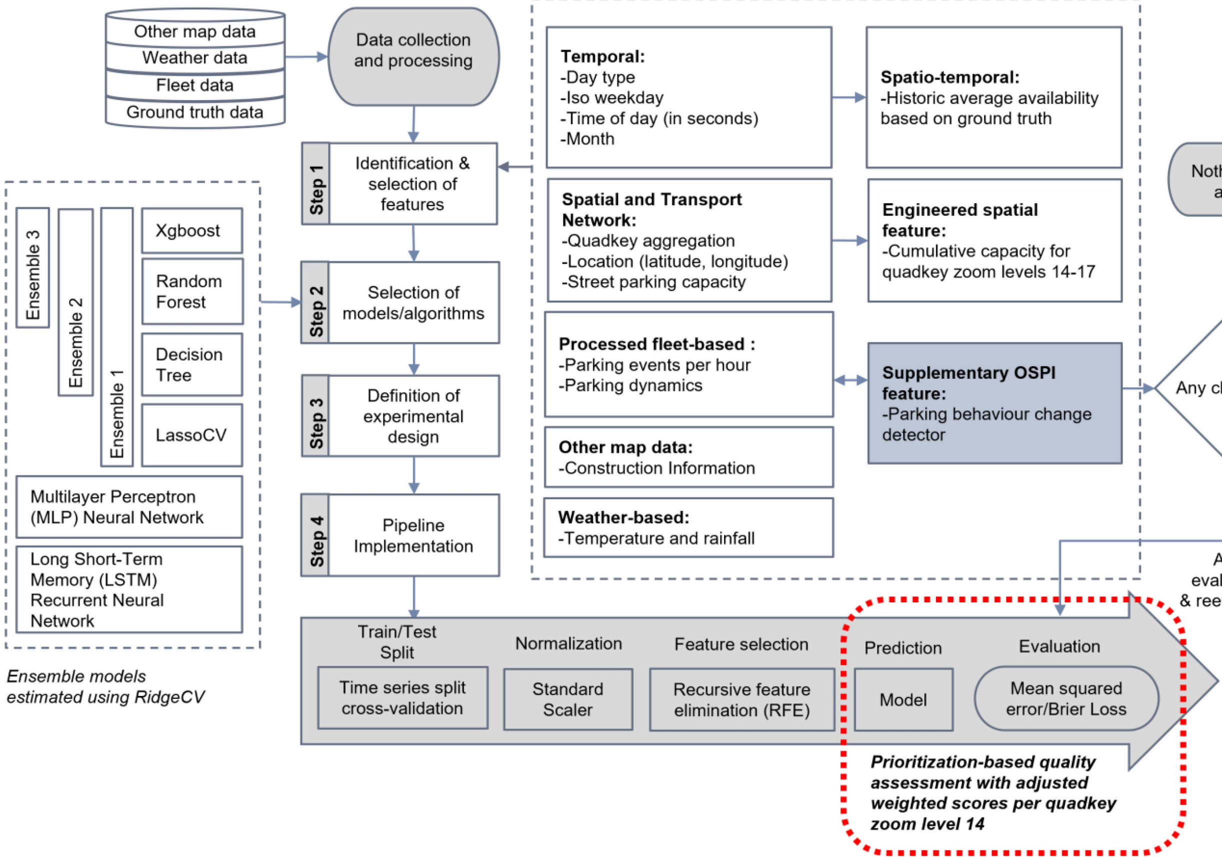


FIGURE 5. Development methodology workflow for an OSPI system.

cannot be directly inferred. Additionally, further aggregate features from basic attributes such as parking capacity were developed as described in Section V.

The OSPI availability prediction models were developed in four main steps (see Fig. 5). The overview of each step is described below. All machine learning implementation besides Xgboost was done using scikit learn [40] in Python.

1) IDENTIFICATION AND SELECTION OF FEATURES

The pre-requisite to start the development was raw data acquisition as described in Section III-A. As the first step, these datasets were used to engineer relevant on-street parking features that are identified based on related literature and domain knowledge. The descriptions of feature content are explained later in Table 1. The features were categorized as follows: temporal, spatial, weather, ground truth historic availability, fleet (parking events) data-based, and other map data.

2) SELECTION OF ALGORITHMS AND ENSEMBLE MODELS

There is a wide range of machine learning models that could be used for parking prediction. The most promising libraries shown in literature are: gradient boosting decision trees like XgBoost [36], Random Forest, and Decision Tree. Deep learning approaches with neural networks have also recently become widely popular butgiven similar performance scores in comparison with the increase in training and processing time [41] it did not seem to be promising. Furthermore, [8] mentions that neural networks perform well with high number of samples to train with like their 12 million records from Melbourne, but with smaller sizes, it may not be feasible. Also, [6] describes that neural networks are suitable when relationships are unknown and high volume of data is available. In this case, since many studies have explored which features could possibly influence the model, unknown relationships are not a big concern. Nonetheless, two neural network models namely Feed Forward Multilayer Perceptron (MLP) and Long Short-Term Memory (LSTM) were implemented for baseline comparisons of all popular models used for parking studies. This is on top of the following four most popular models that were selected and tested amongst each other: Xgboost, Random Forest, Decision Tree, and LassoCV as the baseline linear regression model. Moreover, to get the best of all models, as done in [35], 3 ensemble models were created using RidgeCV as the final estimator that combines the four models (i.e., excluding neural networks) to avoid overfitting on one model.

TABLE 1. Defined feature categories.

Feature category	Description of feature content	Sample values
Temporal	Only time-related features considering aggregations into time intervals in different time scales and categorization of special days: months, weeks, days, hours, minutes, seconds, weekdays, weekends, holidays, etc.	Months: 1-12 Weeks: 1-52 Days: 1-31 Seconds in a day: 1-86400 Holidays: 0 or 1
Spatial and Transport Network	GPS location, on-street parking capacity features divided or aggregated on different spatial levels	Latitude: 48.138393 Longitude: 11.570882 Capacity per street segment: 20 Capacity aggregated on level 14: 74
Weather	Rain and temperature open data	Rainfall: 11.3mm Temperature: 13 deg Celsius
Historic parking availability	Aggregation of historic parking availability on different tile levels and time intervals (e.g. moving averages) in the past.	Available: 1 Occupied: 0 Moving average of availability on level 14: 0.61
Parking events-based	Automated aggregation in various time intervals of TTPD [38] that describe on-street parking activity on tile zoom level 14, and aggregation in various time intervals (e.g. real-time and moving averages) and tile levels of parked-in and parked-out events; anomalies detected based on the developed behavior change detection (see Section IV)	Parked-in volume aggregated on level 14 on 15-minute intervals: 13 Parked-out volume aggregated on level 14 on 15-minute intervals: 12

3) DEFINITION OF THE EXPERIMENTAL DESIGN (ED) SETUP

An experimental design setup was created to organize the process of evaluating the performance of each model by gradually adding feature categories and changing to different types of machine learning algorithms. The aim of the ED setup is to recreate and identify the best combination across algorithms and data types to allow comparison between the different setups that normally exist in the industry given the available dataset in this study. The industry replica model is developed to the best knowledge of the authors since the actual models cannot be used in publications. This ED also allows to identify if a certain setup mainly reliant on parking events-based features can be on par with an industry model and replace it or outperform the industry-level model. In total 54 ED setups were created as displayed and discussed later in Table 3. Further combinations of features for the experimental design were not necessary since even if more features are added to a prediction model, these are reduced in the feature selection step in the pipeline implementation described next.

4) MODEL PIPELINE IMPLEMENTATION

After setting up the input needed into each model, the next step was to create a pipeline implementation to maintain consistency from data transformation to evaluation. The implementation was done through the following pipeline (see Fig. 5): (1) defined the train and test strategy using the time series split cross-validation (see Fig. 4); (2) features were independently normalized using standard scaler from scikit learn; (3) since a large number of features were created, feature selection was employed using recursive feature elimination (RFE) to recursively reduce the number of features used in a model and eliminate irrelevant input features that either do not help the prediction or are redundant; (4) once the optimal features are selected to make the best predictions, these are passed on to a selected model algorithm, and the hyperparameters are tuned. The parking availability predictions are made to the resolution of a second based on the time of request. When the results are integrated into a system, they conform to the user interface (UI), e.g., to be stable, not change frequently, and update every 5 minutes for example, similar to traffic variable message signs (VMS). (5) The last step is to do the evaluation using a metric. Hours that have no ground truth data are excluded from evaluation and are a limitation of this study. Nonetheless, these hours are also considered unimportant hours in Munich based on the study of Gomari et al. [42]. The selected metric for analysis in this study was the Mean Squared Error (MSE) as described below, which is also called the Brier Loss for cases with binary outcomes:

$$MSE = \frac{1}{N} = \sum_{t=1}^{N} (p_t - 0_t)^2 \tag{1}$$

where p is the predicted probability outcome, o is the observation at instance t (0 means there was no available parking spot, 1 means there was at least one available spot), and N is the total number of instances.

MSE is used here as it can punish probability predictions that are farther away from the binary observed ground truth. For further insights, additional metric scores are calculated using the Mean Absolute Error (MAE) and the Root Mean Square Error (RMSE) which can be found in the Appendix. Additionally, given the BMW user-centric system in this study, the proposed prioritization-based quality assessment of [42] is implemented. This method essentially adjusts the scores by taking the weighted sum of the scores of each quadkey at zoom level 14, denoted as KPI_p. The importance weights are based on the total volume of parking events recorded per quadkey over last 3 months of the study period.

$$KPI_P = \sum_{s=1}^{N} KPI_q \times w_q \tag{2}$$

$$w_q = \frac{PEVolume_q}{\sum_{q=1}^{N} PEVolume_q} \tag{3}$$

where KPI_q is the KPI of a quadkey, w is the importance weight assigned to a quadkey, and $PEVolume_q$ is the parking events volume in a quadkey.

All data science tasks carried out in this paper were performed in the Python scripting language. The main packages used were as follows: ADTK, xgboost, Pandas, GeoPandas, Numpy, OSMnx, Matplotlib, Seaborn, Statsmodel, PySal, Scikit-learn, and PyTorch.

IV. SUPPLEMENTAL DYNAMIC OSPI SYSTEM FEATURE: PARKING BEHAVIOR CHANGE DETECTION (PBCD)

An on-street parking availability prediction model is the core component of an OSPI system. This section presents an added-value component and feature to an OSPI system (see Fig. 5) that provides additional dynamicity external of a prediction algorithm, but still part of the OSPI system. The availability of parking events data provided the opportunity to develop a parking behavior change detection (PBCD) model to enhance a user's experience. The PBCD model described here was mainly developed to detect static longer-term changes. Long term is defined as changes that remain in place for at least some defined duration of days ranging from 3 days to 2 weeks. The idea is that the detector allows flagging of potential anomalies due to parking behavioral changes in a city's neighborhood. This then allows an update in the availability predictions made and change the values to zero to represent unavailable spots. Mainly detected are street parking capacity changes or parking rule changes that impact an OSPI system's performance. Such an automatic fleet-based change detection system aims to keep on-street parking maps up-to-date. Early detection of impactful changes helps keep the parking map reliable, accurate, and reduce costs. Furthermore, a PBCD system can alert evaluators to assess the quality of their OSPI models in identified areas by the detector.

The following sections describe the development process and the evaluation carried out for partial validation of the detector. An extensive analysis of the PBCD model is not within the scope of this study. In this paper, only the current status and potential of a PBCD model as an added component within an OSPI system is discussed.

A. METHODOLOGICAL FRAMEWORK FOR THE PARKING BEHAVIOR CHANGE DETECTION (PBCD) MODEL DEVELOPMENT

The complete workflow for the PBCD is illustrated in Fig. 6. The first step after importing parking events fleet data and the on-street parking network was to filter out and process the data. Minimum spatial level and data volume requirements were set to enable behavior change detection. Initially, the spatial requirement heuristically was set to a sub-street quadkey level 17 (approximately 306m x 306m). Each sub-street could contain more than one street block (i.e., intersection to intersection). A sub-street level analysis was chosen instead of street or block level since it was observed that disruptions only occur in small portions of a street affecting only a few parking spots. To minimize noise in the change detection, only sub-street quadkeys at level 17 with parking events greater than 100 for the whole duration of study are chosen for analysis to lessen ambiguity in results. Next, after

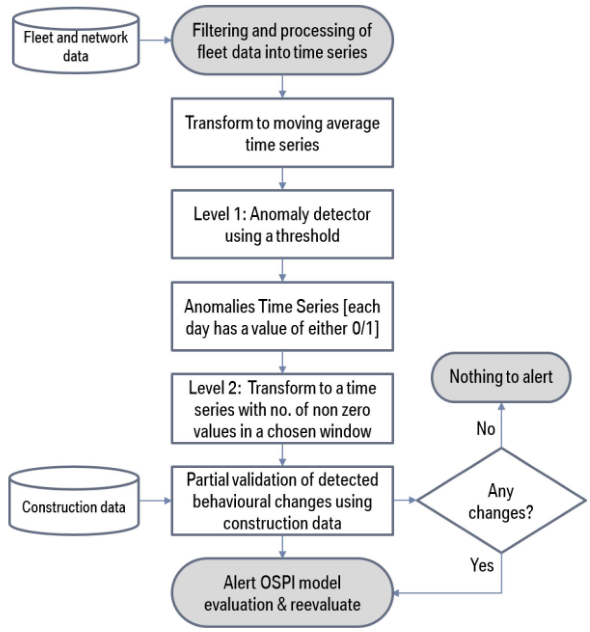


FIGURE 6. Methodology workflow for developing a parking behavior change detection model.

processing the data is converted into a time series for each sub-street.

The rule-based anomaly detection model developed was executed as a **two-level model** shown in Fig. 6 below. A rule-based approach was chosen heuristically based on the known disruptions in the city. For level 1, a threshold was set, and each day with a daily on-street parking volume below this was considered as anomaly and labelled as 1 (with anomaly) or 0 (no anomaly). For an anomaly to be qualified, it must satisfy the level 2 condition, which was done using a rolling aggregator that sums up anomalies and behaves consistently over a defined window number of days based on an experimental design.

The **level 1** detection: a moving average with a window size of 7 days was chosen heuristically for smoothing and transforming the time series. This transformed time series was then used to identify the first level behavioral anomalies. To further eliminate ambiguity, the removal of holidays and weekends before level 1 detection was done, to remove drops on these days, but nonetheless, nothing changed in terms of anomaly detection, indicating that these days do not impact the model. The main factor in the level 1 detection is the testing of different threshold values as cut off values. All the days in the time series which had fewer parking events than the respective threshold value were considered as anomalies. For instance, given the set threshold at 10%, all the days in the time series with parking events less than 10% quantile value are anomalies. This method ensures that all the days with comparatively fewer activities reported are identified as potential longer-term anomalies and can be marked for further analysis.

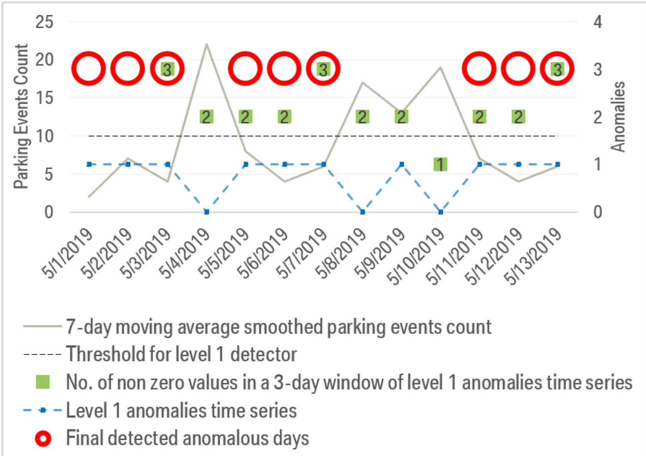


FIGURE 7. **Anomaly detection model instance for time window of 3 days.**

As an input to *level 2* after the threshold detector, each day in the generated anomalies time series was classified as either having a value of 0 (not an anomaly) or 1 (anomaly). Thereafter, the level 2 detector transformed the time series by performing a rolling aggregate to identify the number of non-zero values, i.e., number of anomalies of level 1 for a defined window size in days. If all the values in the considered window are 1, then all are considered as second level anomalies. Even if one of the values in the window is 0, which is not an anomaly, all the remaining values are also considered as 0. For instance, when a window of 5 days is considered, if all the days in that window are first stage anomalies, then all of them are also second stage anomalies; however, if even one day of the 5-day window is not a first stage anomaly, then all the 5 days are dropped as potential anomalous behavior. If both levels are satisfied, a warning can be triggered to change the availability status after 5 days regarding drop in the parking activity of the sub-street, which can be flagged due to a disruptive activity, such as construction, rule change, or some special event.

Fig. 7 illustrates an abstract example of the level 1 and level 2 detections from the PBCD model. The solid line represents the imaginary sample of parking events time series data after performing a 7-day moving average. Now, considering 10 parking events counts as the threshold value (dotted black line), all the days with park event values less than 10 are anomalies after level 1. This new time series with values 0 or 1 is plotted as the dashed blue line. Considering a time window of 3 days for level 2, the green squares show the values (count of number of 1's in the 3 days window) after level 2, which can be 0 or 1 or 2 or 3 and the red circles are the final anomalies after level 2, which are considered as final potential parking behavior changes. These are obtained by considering the green circles with count equal to 3 and two respective previous days as final anomalies.

The percent anomalies omitted after level 2, left-over anomalies, i.e., the days which turned out to be anomalies after level 1 but are rejected in level 2 (the day 2019-05-09 in the Fig. 7) are considered as omitted anomalies and these could also be due to construction (see Fig. 8). For instance, if the considered window in the level 2 is 15 days and all the 14 days in a window are anomalies after level 1. After level 2, none of the dates in that window are considered as anomalies as they do not satisfy the criteria of level 2. But still, they could be due to construction and therefore it is important to capture the percent of omitted anomalies which could be potential anomalies. Percent omitted anomalies within the days where there is a construction event reported could have more chances of becoming an anomaly and therefore these are also evaluated separately.

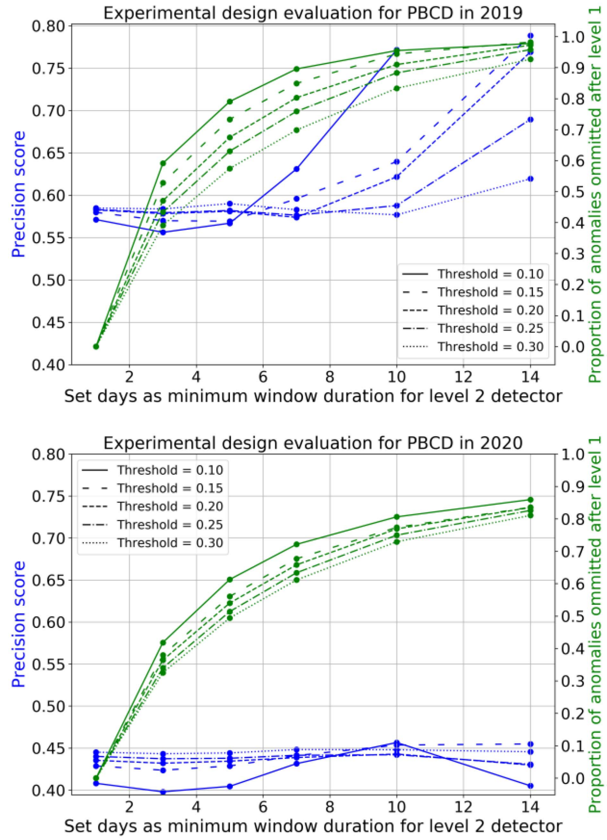


FIGURE 8. **Evaluation precision scores (left y-axis) of each experimental design setup for the parking behavior change detection model including the percentage of anomalies filtered after level 1 detection (on the right y-axis).**

B. EVALUATION EXPERIMENTAL DESIGN OF PBCD MODEL

To evaluate the capabilities of the defined PBCD model, the following experimental design was defined: basically, there were 5 threshold values from below 10% to 30% at 5% intervals, and 6 minimum duration values namely 1, 3, 5, 7, 10, 14, resulting to 30 experimental design setups to check the precision scores. The calculated precision score is a partial validation that presents the percentage of detected parking behavior anomalies that coincide with construction activities, although there may be other reasons for anomalies.

Construction events dataset from HERE Maps was used for the partial cross-validation of the PBCD model. It must be noted that the construction dataset may have a few shortcomings as well, such as: latency in updates and lack of information on impact on parking. The only construction information used were the period of construction and the location or street, where the construction works were observed. Each on-street parking behavior change detected by the model on quadkey level 17 was validated against the existence of construction on street level. If there is a construction on a particular day, then that day is considered as an anomaly. These days with construction events are considered as known anomalies. The precision is defined as:

$$precision = \%\ anomalies\ within\ construction = \frac{TP}{TP + FP} \tag{4}$$

where the observed value is the construction report by HERE maps and detected is an on-street parking behavior change detection. True Positive (TP) is any day which is a model anomaly and a known construction anomaly, and False Positive (FP) is defined as any day which is a detected anomaly but an unknown anomaly.

Based on field inspection, a construction observation does not necessarily mean the on-street parking segment was closed, thus, not all days with construction coincides with an anomaly. It was more often the case that when the road was open, then a parking lane was taken for this, hence, the parking segment was obstructed. Based on the sanity check of construction precision score, which means detecting that for at least one day, an anomaly is recorded within the construction period, we were able to detect at least one disruption in on-street parking for each construction event. The construction sanity precision score of 1.0 for all construction events means that all were detected at some point during their reported period of construction on a specific street. However, the overall anomaly precision scores are lower (see Fig. 8) given that there were identified changes that were not within any construction period. Hence, an anomaly detected by the model is not always caused by construction. Other detections could be other longer-term changes due to parking rules changes, an event occurring at that place for a certain period, or other potential unknown anomalies. It is also possible that, the model anomalies estimated are false change detections. This means not all model anomalies are actual changes but could be because of model inaccuracies.

C. MAIN FINDINGS FROM THE PBCD MODEL

The precision scores corresponding to the various combinations of threshold values for level 1 and the minimum window duration in days for level 2 are presented in Fig. 8. Both the scores for the 2019 and 2020 parking events data are presented. In most cases, it is observed that, higher minimum window duration values for level 2 correlates with a higher precision score. Concurrently, many level 1 anomalies are filtered out as seen with the green lines in the figure. The scores for the 2019 experimental design range from 0.55 to 0.79, while the spread is from 0.39 to 0.46 for the year 2020. The reason for the big difference in precision scores between 2019 and 2020 is the range of data used. In 2019 only 5 months of data from May to September was available, while for 2020 it was 9 months from February until October. Henceforth, the possibilities of detecting more anomalies throughout the year. Another reason for the difference is that anomalies detected in 2020 may not be due to construction; an example could be anomalies from varying restrictions due to the COVID-19 pandemic that started in March 2020 – although this is not tackled here. For both 2019 and 2020, it can also be observed from Fig. 8, that the percent proportion of anomalies omitted after level 1 increases as the minimum window duration is increased. For 2019, the precision improves as more level 1 anomalies are omitted, meaning they are unlikely to be an actual anomaly. However, for 2020, the precision score remained on the same level throughout the different experimental designs as seen in the graph. Similarly, this is attributed to other possible anomalies not related to construction.

Nonetheless, these precision scores are acceptable as it can detect some behavioral changes for which more than 55% and 40% precisions were achieved that are attributed to construction for 2019 and 2020, respectively. This is sufficient as far as the goal to use the PBCD model only as an additive component on top of the availability prediction model (see Section III-A).

Considering all the setups, the most optimal parameters are 0.20 as the threshold for the level 1 detector and the minimum window of 10 days for the rolling aggregator at level 2. With this setup, the parking behavior change detection (PBCD) model developed can detect long term disruptions which last for at least 10 days - anything below this period is neglected. The aim of the developed model was to detect long term static anomalies signaled by the drop of parking events caused by construction, rule change, or a significant infrastructural change, among others. Anomalous activities that increase the number of parking events were not part of this study. In summary, the developed model is valuable and can be used as a trigger functionality in a navigation app to flag potential changes to on-street parking provisions and as an alert for the evaluation of OSPI systems. Furthermore, the feature can be incorporated in the proposed OSPI system described in the next chapter by changing predictions to 0 for unavailability of on-street parking spots.

V. DEVELOPMENT OF A DATA-DRIVEN OSPI SYSTEM

This section presents a comprehensive comparison of different OSPI availability models based on the pipeline implementation discussed in Section III-B4) that can be used as part of the proposed OSPI system. The specific features engineered, elaboration on the usage of each feature category, and the model evaluation are discussed here as well.

A. FEATURES ENGINEERED

A relevant parking prediction study in Munich was carried out by [6] in 2016, wherein they discovered that weekday, location, temperature, and time of the day significantly improve their model performance, while information regarding traffic, holidays and rainfall only had a secondary influence. Hence, apart from traffic information, all the other features were also created and enhanced in this study. In total 102 features were extracted from the raw data available. The breakdown is as follows: 15 time-related, 7 space and location-related, 2 weather-related, 9 based on historic parking availability, 54 features related to parked-in and parked-out events, 12 related to aggregated parking events data called temporal trend of parking dynamics (TTPD) as defined in [38], and 3 related to parking behavior change detection (see Section IV). The description in Table 1 provides more information.

In summary, to create more generalized features, all the data except weather, were aggregated on different quadkey zoom levels; this is a standardized partitioning of the world map into tiles provided by Microsoft's Azure Maps [43]. Aggregation was done from zoom levels 14 corresponding to a tile size of 2446m x 2446m to smaller sizes up to level 17 of 306m x 306m. For the parking events-related and historic parking availability features, different horizons of moving averages slices were tested. A *slice* is a spatio-temporal boundary consisting of a specific quadkey and hour within the 168 hours of the week. These moving averages include taking the average value over the last 2, 4, 6, 8 hours or looking at the same week-hour and quadkey (i.e., slice) over the last 2, 4, 6, 8 calendar weeks. Another averaged value was, for example, taking the average number of parking events at each slice from the last month.

B. MODELS AND TUNED HYPER-PARAMETERS

The optimal hyperparameters of the models change depending on the feature and the nature of the problem tackled. It was observed within all the experimental design setups, the tuned hyperparameters only marginally helped to improve the models relative to the improvements brought by features included in a model. The tuned values displayed in Table 2 are those of experimental design setup 6, which is chosen as the sample setup of the analysis.

The optimal parameters were determined using exhaustive grid search (i.e., GridSearchCV), when it was feasible, and randomized parameter optimization (i.e., RandomizedSearchCV) [40] when model runs take much more time, like in the case of the Random Forest models. For model parameters not listed in Table 2, the default values were taken [36], [40]. The 3 ensemble models created within this paper combines the different standalone models using RidgeCV, which is a linear regression model. The default alpha parameter was taken for the ensemble models.

On average, Xgboost [36] was the best standalone machine learning algorithm tested in this paper. Xgboost is a type of gradient tree boosting system, which is a tree ensemble

TABLE 2. Models and tuned hyperparameters.

Xgboost		**Random Forest**	
Parameter	Value	Parameter	Value
learning_rate	0.04	n_estimators	800
n_estimators	135	min_samples_split	10
max_depth	7	max_depth	110
min_child_weight	5	min_samples_leaf	4
gamma	0	max_features	sqrt'
subsample	0.45	**Decision Tree**	
colsample_bytree	0.65	max_depth	4
objective:	'binary:logistic'	min_samples_leaf	57
scale_pos_weight	1	**LassoCV**	
reg_alpha	0	alpha	0.022
Multilayer Perceptron (MLP) Feed Forward Neural Network		**Long Short-Term Memory (LSTM)**	
hidden_layer_sizes	(21,)	n_epochs	50
activation	'logistic'	num_layers	1
random_state	1	input_size	number of input features
early_stopping	True	hidden_size	21
learning_rate	'adaptive'	learning_rate	0.001

model on its own, wherein the final prediction is based on the prediction values calculated from an aggregation of each tree. The objective function to minimize was set to binary logistic, since the problem dealt with is a logistic regression for binary classification that gives a probability output between 0 and 1. The most important parameter to tune was learning rate; the lower value, the better the predictions had become. After setting a learning rate, the number of trees (n_estimators) is determined. After a certain number of trees, the score does not improve anymore, and it plateaus. For Random Forest, the number of estimators made the most difference, but the scores did not change much in comparison with the default hyperparameters. The biggest difference observed in tuning parameters was with the Decision Tree model. After changing the minimum number of samples to be at a leaf node (min_samples_leaf) from default of 1 to 57, and updating the maximum depth from none to 4, the MSE score improved by 31%. As a baseline example for a linear regression model, LassoCV was used. LassoCV is usually used in regularization in machine learning to avoid overfitting and for feature selection. The only relevant factor to tune here was the complexity parameter alpha which was set to 0.022. For the Multilayer Perceptron (MLP) the hidden layer size was the most relevant. The optimum value for this was the desired number of selected input features after feature selection in the pipeline. For the Long Short-Term Memory baseline model, the number of epochs was the most crucial

TABLE 3. Experimental design and Prioritization-based scores for prediction models.

		Experimental design setup						
	Feature category	**1**	**2**	**3**	**4**	**5**	**6**	
1	**Temporal**	x	x	x	x	x	x	
2	**Spatial**		x	x	x	x	x	
3	**Weather**			x	x	x	x	
4	**Historic parking availability**				x		x	
5	**Parking events-based**					x	x	
	Model (M)	Model Mean Squared Error (MSE) Scores						**Average score**
1	**Xgboost**	0.2468	0.2160	0.2161	0.2166	0.2168	0.2159	0.2214
2	**Random Forest**	0.2467	0.2170	0.2161	0.2174	**0.2152**	0.2165	0.2215
3	**Decision Tree**	0.2398	0.2412	0.2400	0.2418	0.2354	0.2352	0.2389
4	**LassoCV**	0.2408	0.2316	0.2291	0.2251	0.2294	0.2253	0.2302
5	**Ensemble 1 = M1+M2+M3+M4**	0.2387	0.2154	0.2144	0.2152	0.2148	0.2157	0.2190
6	**Ensemble 2 = M1+M2+M3**	0.2423	0.2163	0.2150	0.2174	0.2148	0.2165	0.2204
7	**Ensemble 3 = M1+M2**	0.2447	0.2154	0.2148	0.2167	**0.2146**	0.2162	0.2204
8	**MLP Neural Network**	0.2330	0.2245	0.2239	0.2288	0.2288	0.2245	0.2273
9	**LSTM RNN**	0.2385	0.2392	0.2402	0.2448	0.2422	0.2437	0.2414
	Average score	0.2413	0.2241	0.2233	0.2249	0.2236	0.2233	
	Legend for each MSE score	Low	High					
	Legend for average MSE score	Low	High					

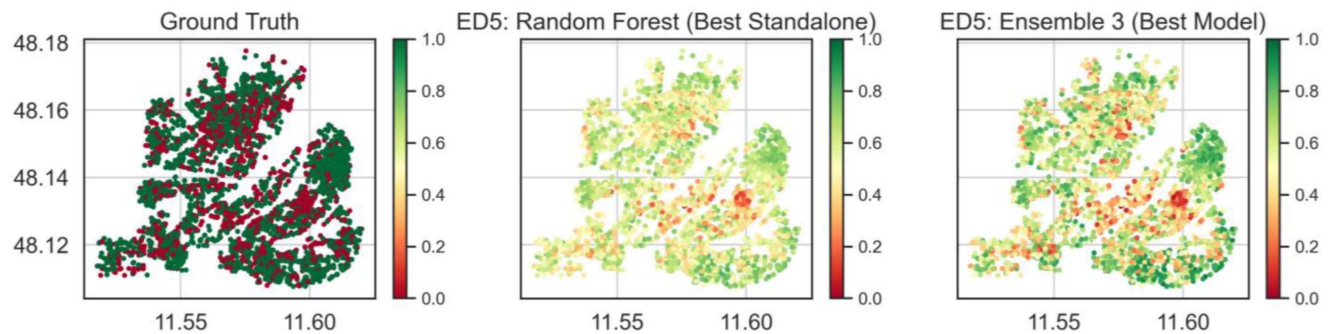


FIGURE 9. Ground truth observations (left), test set predicted probability maps of the best standalone model (middle) and the best overall model (right).

to optimize training time. After 50 epochs, the score was not improving anymore.

The main finding in the hyperparameter tuning task for the prediction model of the OSPI system is that the features selected and passed on to a model are more important compared to hyperparameter optimization unless a new algorithm is to be developed. However, for a simpler model like Decision Tree, the parameter values have a larger impact on the evaluation score. Nonetheless, tuning is vital in maximizing the performance of prediction algorithms used.

C. EXPERIMENTAL DESIGN AND EVALUATION OF OSPI AVAILABILITY MODELS

The comparative analysis of the various models based on the experimental designs is discussed in this section: the mean MSE scores, the features that help a model, the features that can replace other ones, geographical analysis, and the performance of different algorithms. The systematic process of evaluation was defined through several experimental design (ED) setups as described in Table 3. Different feature categories were gradually added as part of the experiment. For each of the 6 EDs, 9 models were used, totaling to 54 setups.

The calculated prioritization-based MSE scores [42] are illustrated in Table 3. The worst performing model scores are achieved at ED1 when only temporal features are considered. In this scenario, it can be observed that the neural network models outperform the other models as they are more capable of finding latent variances that the other algorithms cannot determine without more features. The best performing model among the 54 setups was Ensemble 3 at ED5 with a 0.2146 score, which combines Xgboost and Random Forest while taking all features except historic parking availability-based features. The best standalone model is also at ED5: Random Forest with a score of 0.2152. Each predicted probability from the test set of around 7000 observations is mapped in Fig. 9. To demonstrate the sensitivity to time in terms of average parking availability, see Fig. 10. This is the average parking availability versus the average parking probability prediction based on the best model per week-hour. And as seen, the model can predict in line with the availability patterns. If there are discrepancies, these are

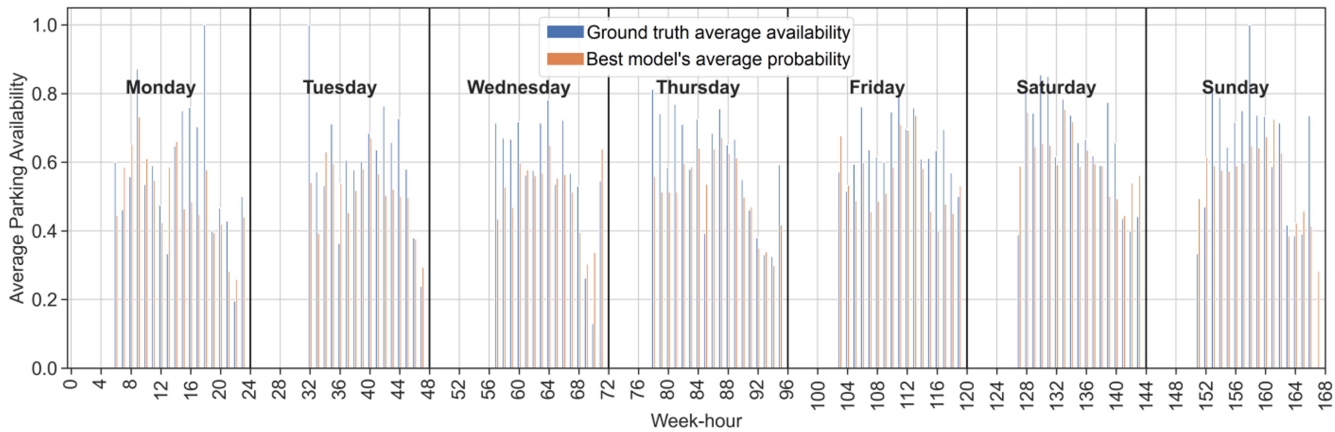


FIGURE 10. Average parking availability based on ground truth versus the best prediction model's average parking probability.

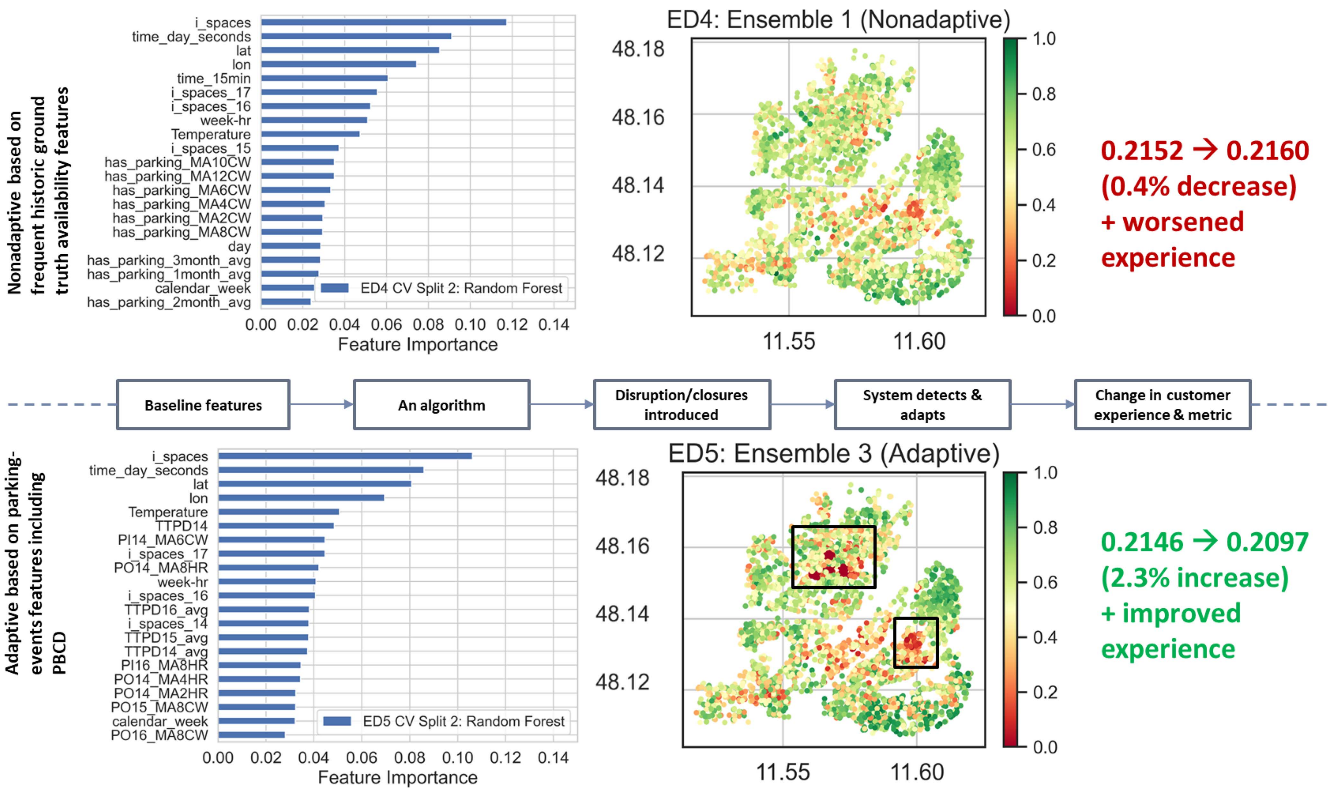


FIGURE 11. The stages of comparison evaluation and the differences between a nonadaptive (top) and an adaptive (bottom) OSPI system.

considered in the prioritization-based scores, which adjusts in accordance with spatio-temporal importance [42].

Fig. 9 illustrates the difference between the predictions made between the two models. ED5: Ensemble 3 has a wider spread of prediction, meaning the spread is farther from the mean. This can be observed in the maps by the larger contrast in color in the best model's predicted probability map. This translates to the model making more confident prediction.

An objective of this study was to reduce reliability on ground truth data collection and have a more dynamic data-driven OSPI that does not rely on continuous ground truth collection to reduce costs. There are **two main stages** to assess this: (1) see if an alternative model, in this case, the parking events-based model (ED5) is on par with existing industry-level models as represented by ED4; and (2) illustrate the dynamicity and advantage of the alternative model. The stages of comparison are demonstrated in Fig. 11.

Stage 1: In Table 3, it is shown that the performance of ED5 across the different algorithms implemented is in most cases outperforming the ED4 models. This makes it clear that ED5 can be a feasible alternative to an industry model that focuses on historic parking availability features for its dynamicity (see Fig. 11). To compare features, specifically, the industry-level model at ED4 using Random Forest can

TABLE 4. Prioritization-based scores after introducing disruption/closure in 5 out of 772 street segments.

	Model (M)	Model Mean Squared Error (MSE) Scores						Average score
	Experimental design (ED) setup	**1**	**2**	**3**	**4**	**5**	**6**	
1	**Xgboost**	0.2483	0.2172	0.2170	0.2180	0.2120	0.2107	0.2205
2	**Random Forest**	0.2480.	0.2181	0.2168	0.2187	**0.2103**	0.2116	0.2206
3	**Decision Tree**	0.2418	0.2425	0.2415	0.2431	0.2301	0.2299	0.2382
4	**LassoCV**	0.2410	0.2302	0.2277	0.2244	0.2242	0.2201	0.2279
5	**Ensemble 1 = M1+M2+M3+M4**	0.2395	0.2163	0.2150	0.2160	0.2099	0.2106	0.2179
6	**Ensemble 2 = M1+M2+M3**	0.2426	0.2173	0.2157	0.2184	0.2099	0.2114	0.2192
7	**Ensemble 3 = M1+M2**	0.2448	0.2164	0.2155	0.2177	**0.2097**	0.2112	0.2192
8	**MLP Neural Network**	0.2349	0.2232	0.2227	0.2295	0.2237	0.2195	0.2256
9	**LSTM RNN**	0.2408	0.2411	0.2421	0.2466	0.2366	0.2382	0.2409
	Average score	0.2424	0.2247	0.2238	0.2258	0.2185	0.2181	
	Legend for each MSE score	Low	High					
	Legend for average MSE score	Low	High					

be compared to the best standalone model ED5: Random Forest. The reason standalone models are compared is that feature importance can be directly extracted as opposed to an Ensemble model. This is obtained from the built-in feature importance attribute, determined by the proportion of the number of times a feature appeared in a tree by a model. The optimal number of features selected through various trials was 21. Thus, whenever more features were available, the 21 best features that best generalize the parking prediction were selected. The differences in features used and the respective importance factors are shown in Fig. 11. The most important features are the primary spatial and temporal features: parking spaces or capacity, time of day in seconds, and GPS location. Looking at Table 3, the primary features support each other. There are variances only captured by spatial features, that significantly improve the performance that are not captured by temporal features as seen in scores of ED1. In the ED5: Random Forest feature importance graph (lower left in Fig. 11), it can be seen that 11 out of 21 features are parking event-based. Looking at Table 3, ED5: Random Forest attains a score of 0.2152, while ED4: Random Forest attains 0.2174. This presents a 1% difference in score and can be concluded that ED5: Random Forest after replacing historic parking availability-based features with parking events-based features does not impact the performance. Thus, for stage 1 of the assessment, it can be an alternative to an industry model.

Furthermore, from the comparison of features it was discovered that aggregated spatial features appear to capture variances previously unknown. This is beneficial to further reduce reliance on historic parking availability features. On top of on-street parking capacity on a street-level, denoted as i_spaces in Fig. 11, aggregation of capacity on level 14, 16, and 17, labelled as i_spaces_14, i_spaces_16, and i_spaces_17, respectively, are capable of capturing variances and better generalize. To the best knowledge of the authors, this is a new finding that has not been discussed in research, as majority focus on directly using street parking capacity on a street-level, when this data is available. This static feature can also be updated with a dynamic feature such as PBCD that detected disruptions as discussed next.

Stage 2: For the next stage, the dynamicity is important, hence, as shown in Fig. 11, the best models are used for score comparison, and these are ED4: Ensemble 1 and ED5: Ensemble 3, respectively. To explicitly demonstrate the dynamicity of the parking events-based models at ED5 with the integration of a PBCD, on-street parking disruption or closures were artificially introduced to 5 of 772 street segments in the study area. For the entire study period, the ground truth availability is then changed to zero. This was to illustrate the difference in the performance scores for models that detect these anomalies and adapt. As seen in Fig. 11 in the two predicted probability maps, the adaptive OSPI system using parking events features and PBCD can detect the closures that are denoted with the boxes. This is visibly not detected in the nonadaptive model of ED4: Ensemble 1. Before disruption, the scores are quite similar with the nonadaptive model scoring 0.2152, and the adaptive model scoring 0.2146. However, after the closure, only the adaptive one improves its score as it is able to change its predictions based on a trigger from its PBCD. In large cities, these disruptions are difficult to detect. And often in a city like Munich, a parking closure that is left unnoticed and not updated in the system causes a compounding effect on parking search that can lead to a worsened experience of the OSPI system. Thus, a system that relies on parking events and its added PBCD feature does not only lessen the dependence on manual ground truth observations to check for disruptions, it also automatically improves user experience of the proposed OSPI system.

The complete changes in scores are shown in Table 4 with the updated scores after the introduction of disruption and Table 5 Shows the percentage difference in comparison with the prioritization-based MSE scores in Table 3. To summarize, the spread of scores based on feature category experimental design and by model used is shown in Fig. 12 Based on the average scores per feature category ED, ED 6 scores the best with an average MSE of 0.2181 after introduction of disruption, followed by ED 5, 3, 2, 4, and 1 (see Fig. 12). It is also apparent that the adaptive models in ED5 and ED6 outperform the other 4 EDs proving their advantage over models that need manual ground truth

TABLE 5. Score percentage difference between after and before disruption per model and experimental design.

	Model (M)	Model Mean Squared Error (MSE) Scores						Average score
1	**Xgboost**	-0.6%	-0.6%	-0.4%	-0.6%	2.2%	2.4%	0.4%
2	**Random Forest**	-0.5%	-0.5%	-0.3%	-0.6%	2.3%	2.3%	0.4%
3	**Decision Tree**	-0.8%	-0.5%	-0.6%	-0.5%	2.3%	2.3%	0.3%
4	**LassoCV**	-0.1%	0.6%	0.6%	0.3%	2.3%	2.3%	1.0%
5	**Ensemble 1 = M1+M2+M3+M4**	-0.3%	-0.4%	-0.3%	-0.4%	2.3%	2.4%	0.5%
6	**Ensemble 2 = M1+M2+M3**	-0.1%	-0.5%	-0.3%	-0.5%	2.3%	2.4%	0.5%
7	**Ensemble 3 = M1+M2**	0.0%	-0.5%	-0.3%	-0.5%	2.3%	2.3%	0.6%
8	**MLP Neural Network**	-0.8%	0.6%	0.5%	-0.3%	2.2%	2.2%	0.7%
9	**LSTM RNN**	-1.0%	-0.8%	-0.8%	-0.7%	2.3%	2.3%	0.2%
	Average score	-0.4%	-0.3%	-0.2%	-0.4%	2.3%	2.3%	

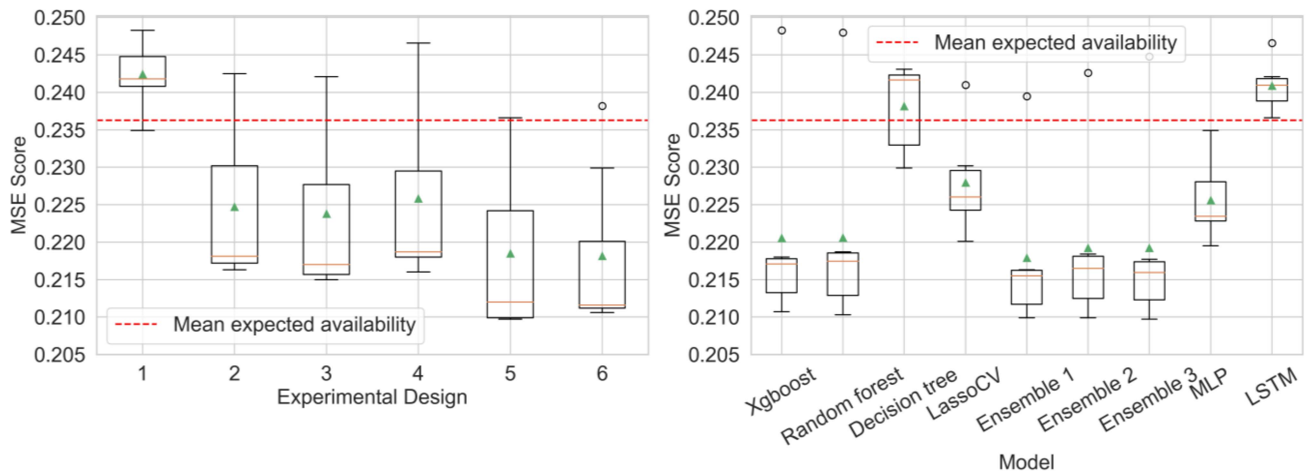


FIGURE 12. Spread of scores after disruption. Average cross-validation MSE score box plot by feature category experimental design (ED) and average cross-validation MSE score box plot by machine learning algorithm/model.

observations to identify disruptions. Meanwhile, the worst score is recorded when using Random Forest with only temporal features (ED 1). However, when spatial features are added, the Random Forest model significantly improves its performance. The same is also observed with Xgboost and the ensemble models. The Decision Tree and LassoCV models only improve with the gradual increase in features. For MLP the trend is unclear as a decline in performance was observed at ED4. For LSTM, since it is a model that primarily relies on organized time-series data, there is no significant improvements after the temporal features introduced at ED1. This proves that a domain knowledge driven models that rely on feature engineering can outperform baseline neural networks as those presented here.

ED4 includes historic parking availability features, but comparatively, it performed worse than the previous step on average. This changes at ED6 when parking events-based features are added, resulting to the best average MSE. Comparing ED5 against ED4, it can be concluded that in most models, the parking events-based features help more than the historic parking availability features. The more features provided, on average, models can capture more variances to make adjustments necessary to improve predictions – this is even more apparent with LassoCV, a simple linear model. The boxplot for the average MSE for LassoCV in Fig. 12 shows the shift from 0.2410 at ED 1 to 0.2201 at ED6 that is tabulated in Table 3.

D. DISCUSSION ON THE LATENCY OF THE PROPOSED OSPI SYSTEM

Fig. 4 shows that training only needs to take place every three months as conducted in this study. This means, the estimator factors in the machine learning algorithms employed remain the same. These factors are calibrated and adjusted based on the input values in the training sets. Table 2 presents the feature categories that contain different engineered features for each. Each feature takes a different value input. The temporal features get input in relation to a timestamp of a request. The spatial features are static based on the parking map but can be updated when the parking behavior change detection (PBCD) feature detects long-term closures or disruptions in capacity. However, for dynamic feature categories such as weather, historic parking availability and parking events-based features the system relies on ingested data. The feed or ingestion rate is different for each. For weather, hourly temperature and rainfall data can be captured. For ground truth historic parking availability data, this can only be fed into the system at random intervals depending on when data collection is scheduled with on-site observers. Thus, the system takes the historic averages

TABLE 6. Experimental design and prioritization-based root mean squared error scores for prediction models.

		Experimental design setup						
	Feature category	**1**	**2**	**3**	**4**	**5**	**6**	
1	**Temporal**	x	x	x	x	x	x	
2	**Spatial**		x	x	x	x	x	
3	**Weather**			x	x	x	x	
4	**Historic parking availability**				x		x	
5	**Parking events-based**					x	x	
	Model (M)	Model Root Mean Squared Error (RMSE) Scores						**Average score**
1	**Xgboost**	0.4965	0.4640	0.4642	0.4649	0.4649	0.4639	0.4697
2	**Random Forest**	0.4964	0.4654	0.4644	0.4660	**0.4634**	0.4648	0.4701
3	**Decision Tree**	0.4894	0.4905	0.4894	0.4913	0.4847	0.4845	0.4883
4	**LassoCV**	0.4907	0.4811	0.4785	0.4742	0.4789	0.4744	0.4796
5	**Ensemble 1 = M1+M2+M3+M4**	0.4885	0.4635	0.4624	0.4635	0.4626	0.4637	0.4674
6	**Ensemble 2 = M1+M2+M3**	0.4922	0.4644	0.4631	0.4658	0.4627	0.4646	0.4688
7	**Ensemble 3 = M1+M2**	0.4946	0.4635	0.4629	0.4652	**0.4624**	0.4643	0.4688
8	**MLP Neural Network**	0.4826	0.4734	0.4727	0.4776	0.4779	0.4735	0.4763
9	**LSTM RNN**	0.4879	0.4887	0.4898	0.4945	0.4919	0.4935	0.4911
	Average score	0.4910	0.4727	0.4719	0.4737	0.4722	0.4719	
	Legend for each MSE score	Low	High					
	Legend for average MSE score	Low	High					

available from the last collection period. This is also the reason that manual collection is not deemed feasible. Parking events-based features are engineered to aggregate values for several intervals with the shortest being 15 minutes. This means, the OSPI system predicts parking with a 15-minute latency or 15 minutes into the future. Thus, if information is requested now, the parking events-based features feed the aggregated value in the last 15-minute interval.

In the occurrence of feed failure errors, the system reverts to the last historic averages to fill in the missing data. Detailed feed failures with regards to parking events cannot be covered in this study due to the lack of access to relevant data at the point of collection.

E. MAIN FINDINGS AND DEPLOYMENT OF A DATA-DRIVEN OSPI SYSTEM

The best model based on the analysis is ED5: Ensemble 3 using temporal, spatial, weather, and parking events-based features. The ensemble model is a combination of Random Forest and Xgboost. The ED5: Random Forest standalone model was the best performing. Once combined with Xgboost, the resulting model was able to take the best of the two algorithms by learning from the weak predictions and replacing them with the advantages of the other. This is illustrated by the starker difference in predicted probabilities of the ED5: Ensemble 3 model as shown in Fig. 9. This is also interpreted as a more confident prediction model since the values are closer to a binary outcome, while improving the prioritization-based MSE score performance.

Even though an industrially accepted model such as ED4: Ensemble 1, which mainly relies on manual ground truth for updates and disruption information, model ED5: Ensemble 3 is a better model of choice for companies or institutions that have direct access to reliable incoming fleet data. This is because the best model employs features that rely on continuously available parking events data capable of capturing real-time and up-to-date variances that are needed to adjust the parking availability model. Furthermore, a parking behavior change detection (PBCD) feature based on the parking events improves the performance of the system by detecting disruptions and closures of on-street parking spaces. Such a system reduces the need to send manual observers to collect data to update the system and its relevant associated parking maps.

VI. CONCLUSION AND RECOMMENDATIONS

In the industry, manual data collection is still prevalent to ensure quality. The authors have proposed an on-street parking information system with a parking availability prediction model and a supplementary additive component that provides on-street parking behavior change detection (PBCD) using the parking events dataset. The parking availability prediction model utilizes parking events-based features and enhanced spatial features that have a better capability to generalize on-street parking capacity on different spatial aggregation quadkey zoom levels. The developed parking availability prediction model and methodology can be a competitive alternative to existing models which mainly rely on historic ground truth observations converting it to parking availability features and do not have many adaptive and dynamic features such as the parking events-based ones introduced in this paper. A wide range of feature categories and machine learning algorithms were tested as part of an experimental design to identify the best configuration of features engineered based on domain knowledge and existing algorithms.

One main advantage of the presented approach for a city like Munich, where there is abundant parking events data, is the opportunity to reduce the frequency of ground truth collection since the model can rely on incoming parking events

TABLE 7. Experimental design and prioritization-based mean absolute error scores for prediction models.

		Experimental design setup						
	Feature category	**1**	**2**	**3**	**4**	**5**	**6**	
1	**Temporal**	x	x	x	x	x	x	
2	**Spatial**		x	x	x	x	x	
3	**Weather**			x	x	x	x	
4	**Historic parking availability**				x		x	
5	**Parking events-based**					x	x	
	Model (M)	Model Absolute Error (MAE) Scores						**Average score**
1	**Xgboost**	0.4740	0.4243	0.4283	0.4288	0.4257	0.4248	0.4343
2	**Random Forest**	0.4723	0.4414	0.4437	0.4472	**0.4424**	0.4436	0.4484
3	**Decision Tree**	0.4653	0.4584	0.4602	0.4610	0.4567	0.4558	0.4596
4	**LassoCV**	0.4862	0.4707	0.4687	0.4607	0.4690	0.4613	0.4694
5	**Ensemble 1 = M1+M2+M3+M4**	0.4799	0.4308	0.4296	0.4359	0.4242	0.4274	0.4380
6	**Ensemble 2 = M1+M2+M3**	0.4872	0.4308	0.4299	0.4389	0.4242	0.4284	0.4399
7	**Ensemble 3 = M1+M2**	0.4900	0.4306	0.4304	0.4388	**0.4240**	0.4285	0.4404
8	**MLP Neural Network**	0.4689	0.4494	0.4476	0.4365	0.4376	0.4310	0.4452
9	**LSTM RNN**	0.4703	0.4759	0.4756	0.4785	0.4817	0.4884	0.4784
	Average score	0.4771	0.4458	0.4460	0.4474	0.4428	0.4432	
	Legend for each MSE score	Low	High					
	Legend for average MSE score	Low	High					

data from vehicles. This was proven by the performance ED 5: Ensemble 3 model. Although the model performs well and adapts to disruptions and closures, normal routine ground truth checks are still necessary at intermittent periods. The introduced methodology in this paper however is also limited based on the accessibility of institutions to reliable fleet data that can be used.

It is known that many special events, construction activities, rule changes occur unannounced and undocumented for, hence, the PBCD model presented in this paper can be recommended as an automated flagging component in future OSPI services, that would request for user feedback and confirmation on parking availability. This in return enables faster update of parking maps, while enhancing user experience. It would be interesting to further validate the parking behavior change detection with other data sources such as special events, and rule and infrastructure change data, among others.

The model of interest parking availability prediction model developed in this paper used the following features: temporal, spatial (location and parking capacity spatial aggregates), weather, and parking events-based. Reference [5] demonstrated the value of using their Baidu maps with refined POI data for example. However, in this research it was difficult to obtain reliable POI data without much categorization. Existing open-source POI data are unbalanced and skewed towards restaurants. Future researchers can work on OpenStreetMap POI data with an extensive category definition and cleansing that could be useful for comprehensive development of models in specific cities. The level of OSM POI data coverage is different for each city. Another recommendation is to investigate and evaluate the scores based on priority or important areas in a city [42].

In the future fast processing of videos and images will change the game, but for the meantime, the data volume of parking events is much smaller, and it will remain as a possible source for validating future researches.

APPENDIX

See Tables 6 and 7.

ACKNOWLEDGMENT

The authors would like to acknowledge the support of INRIX in gathering ground truth data and HERE Maps for the provision of construction data, which were both fundamental assets used in this research.

REFERENCES

[1] T. Friedrich et al., “Routing for on-street parking search using probabilistic data,” *AI Commun.*, vol. 32, no. 2, pp. 113–124, 2019, doi: 10.3233/AIC-180574.
[2] K. Gkolias and E. I. Vlahogianni, “Convolutional neural networks for on-street parking space detection in urban networks,” *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 12, pp. 4318–4327, Dec. 2019, doi: 10.1109/TITS.2018.2882439.
[3] H. S. Jomaa, J. Grabocka, and L. Schmidt-Thieme, “A hybrid convolutional approach for parking availability prediction,” in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, 2019, pp. 1–8, doi: 10.1109/IJCNN.2019.8852400.
[4] A. J. Pel and E. Chaniotakis, “Stochastic user equilibrium traffic assignment with equilibrated parking search routes,” *Transp. Res. B, Methodol.*, vol. 101, pp. 123–139, Jul. 2017, doi: 10.1016/j.trb.2017.03.015.
[5] Y. Rong, Z. Xu, R. Yan, and X. Ma, “Du-parking: Spatio-temporal big data tells you realtime parking availability,” in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Disc. Data Min.*, 2018, pp. 646–654.
[6] C. Pflügler, T. Köhn, M. Schreieck, M. Wiesche, and H. Krcmar, “Predicting the availability of parking spaces with publicly available data,” in *Proc. INFORMATIK*, 2016, pp. 361–374.
[7] C. Dosch, M. Farghal, M. Kapsecker, C. Lorenz, and A. Mosharafa, *Parking Prediction in the City of Melbourne*, Technische Universitaet Muenchen, Münich, Germany, 2017.

[8] W. Shao, Y. Zhang, B. Guo, K. Qin, J. Chan, and D. Flora, "Parking availability prediction with long short term memory model," in *Proc. Int. Conf. Green, Pervasive, Cloud Comput. (GPC)*, 2018, pp. 124–137.

[9] S. Yang, W. Ma, X. Pi, and S. Qian, "A deep learning approach to real-time parking occupancy prediction in spatio-temporal networks incorporating multiple spatio-temporal data sources," *Transp. Res. C, Emerg. Technol.*, vol. 107, pp. 248–265, Oct. 2019.

[10] "Maps & map updates." Tomtom Website. Accessed: Aug. 27, 2021. [Online]. Available: https://www.tomtom.com/en_gb/sat-nav/maps-services/map-updates/#mydriveconnect-tab

[11] F. V. Monteiro and P. Ioannou, "On-street parking prediction using real-time data," in *Proc. IEEE 21st Int. Conf. Intell. Transp. Syst. (ITSC)*, 2018, pp. 2478–2483.

[12] F. Shi, D. Wu, Q. Liu, Q. Han, and J. A. Mccann, "ParkCrowd: Reliable crowdsensing for aggregation and dissemination of parking space information," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 11, pp. 4032–4044, Nov. 2019, doi: 10.1109/TITS.2018.2879036.

[13] T. Rajabioun and P. Ioannou, "On-street and off-street parking availability prediction using multivariate spatiotemporal models," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 5, pp. 2913–2924, Oct. 2015, doi: 10.1109/TITS.2015.2428705.

[14] A. Origlia, S. D. Martino, and Y. Attanasio, "On-line filtering of on-street parking data to improve availability predictions," in *Proc. 6th Int. Conf. Models Technol. Intell. Transp. Syst. (MT-ITS)*, 2019, pp. 1–7, doi: 10.1109/MTITS.2019.8883375.

[15] A. Ionita, A. Pomp, M. Cochez, T. Meisen, and S. Decker, "Where to park? Predicting free parking spots in unmonitored city areas," in *Proc. 8th Int. Conf. Web Intell. Min. Semant.*, 2018, pp. 1–12, doi: 10.1145/3227609.3227648.

[16] S. Yang and Z. Qian, "Turning meter transactions data into occupancy and payment behavioral information for on-street parking," *Transp. Res. C, Emerg. Technol.*, vol. 78, pp. 165–182, May 2017, doi: 10.1016/j.trc.2017.02.022.

[17] F. Bock, K. Xia, and M. Sester, "Mapping similarities in temporal parking occupancy behavior based on city-wide parking meter data," in *Proc. Int. Cartograph. Assoc. (ICA)*, vol. 1, 2018, pp. 1–5, doi: 10.5194/ica-proc-1-12-2018.

[18] B. Xu, O. Wolfson, J. Yang, L. Stenneth, P. S. Yu, and P. C. Nelson, "Real-time street parking availability estimation," in *Proc. IEEE Int. Conf. Mobile Data Manage.*, vol. 1, 2013, pp. 16–25, doi: 10.1109/MDM.2013.12.

[19] C. Ajeng and T. H. T. Gim, "Analyzing on-street parking duration and demand in a Metropolitan City of a developing country: A case study of Yogyakarta City, Indonesia," *Sustainability*, vol. 10, no. 3, p. 591, 2018, doi: 10.3390/su10030591.

[20] V. Coric and M. Gruteser, "Crowdsensing maps of on-street parking spaces," in *Proc. IEEE Int. Conf. Distrib. Comput. Sens. Syst.*, 2013, pp. 115–122, doi: 10.1109/DCOSS.2013.15.

[21] F. Bock and S. Di Martino, "How many probe vehicles do we need to collect on-street parking information?" in *Proc. 5th IEEE Int. Conf. Models Technol. Intell. Transp. Syst.*, 2017, pp. 538–543, doi: 10.1109/MTITS.2017.8005731.

[22] L. Stenneth, O. Wolfson, B. Xu, and P. S. Yu, "PhonePark: Street parking using mobile phones," in *Proc. IEEE 13th Int. Conf. Mobile Data Manage.*, 2012, pp. 278–279, doi: 10.1109/MDM.2012.76.

[23] M. Arab and T. Nadeem, "MagnoPark—Locating on-street parking spaces using magnetometer-based pedestrians' smartphones," in *Proc. 14th Annu. IEEE Int. Conf. Sens. Commun. Netw.*, 2017, pp. 1–9, doi: 10.1109/SAHCN.2017.7964915.

[24] K. Ivan, T. Marta, M. Kostadin, and T. Dimitar, "Parking availability prediction using traffic data services," in *Proc. Conf. ICT Innovat.*, 2020, pp. 1–15.

[25] C. Badii, P. Nesi, and I. Paoli, "Predicting available parking slots on critical and regular services by exploiting a range of open data," *IEEE Access*, vol. 6, pp. 44059–44071, 2018, doi: 10.1109/ACCESS.2018.2864157.

[26] 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," *Transp. Res. C, Emerg. Technol.*, vol. 77, pp. 67–79, Apr. 2017, doi: 10.1016/j.trc.2017.01.008.

[27] H. Wang, R. Li, X. Wang, and P. Shang, "Effect of on-street parking pricing policies on parking characteristics: A case study of nanning," *Transp. Res. A, Policy Pract.*, vol. 137, pp. 65–78, Jul. 2020. [Online]. Available: https://doi.org/10.1016/j.tra.2020.04.003

[28] J.-E. Navarro-B, M. Gebert, and R. Bielig, "On automatic extraction of on-street parking spaces using park-out events data," in *Proc. IEEE Int. Conf. Omni Layer Intell. Syst. (COINS)*, 2021, pp. 1–7, doi: 10.1109/COINS51742.2021.9524119.

[29] O. Cats, C. Zhang, and A. Nissan, "Survey methodology for measuring parking occupancy: Impacts of an on-street parking pricing scheme in an urban center," *Transp. Policy*, vol. 47, pp. 55–63, Apr. 2016, doi: 10.1016/j.tranpol.2015.12.008.

[30] N. Arora et al., "Hard to park? Estimating parking difficulty at scale," in *Proc. 25th ACM SIGKDD Conf. Knowl. Disc. Data Min. (KDD)*, 2019, pp. 2296–2304.

[31] F. Richter, S. Di Martino, and D. C. Mattfeld, "Temporal and spatial clustering for a parking prediction service," in *Proc. IEEE 26th Int. Conf. Tools Artif. Intell.*, 2014, pp. 278–282, doi: 10.1109/ICTAI.2014.49.

[32] N. Feng, F. Zhang, J. Lin, J. Zhai, and X. Du, "Statistical analysis and prediction of parking behavior," in *Network and Parallel Computing*. Cham, Switzerland: Springer, 2019, doi: 10.1007/978-3-030-30709-7_8.

[33] A. Bhattacharyya, W. Wang, C. Tsang, and C. Amza, "Semantic-aware anomaly detection in real time parking data," in *Proc. IEEE 15th Int. Conf. Ind. Informat.*, 2017, pp. 486–491, doi: 10.1109/INDIN.2017.8104820.

[34] W. Alajali, S. Wen, and W. Zhou, "On-street car parking prediction in smart city: A multi-source data analysis in sensor-cloud environment," in *Security, Privacy, and Anonymity in Computation, Communication, and Storage* (Lecture Notes in Computer Science 10658). Cham, Switzerland: Springer, 2017, pp. 641–652, doi: 10.1007/978-3-319-72395-2_58.

[35] F. M. Awan, Y. Saleem, R. Minerva, and N. Crespi, "A comparative analysis of machine/deep learning models for parking space availability prediction," *Sensors*, vol. 20, no. 1, p. 322, 2020, doi: 10.3390/s20010322.

[36] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. ACM SIGKDD Int. Conf. Knowl. Disc. Data Min.*, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.

[37] Y. Zheng, S. Rajasegarar, C. Leckie, and M. Palaniswami, "Smart car parking: Temporal clustering and anomaly detection in urban car parking," in *Proc. IEEE 9th Int. Conf. Intell. Sens., Sens. Netw. Inf. Process.*, 2014, pp. 21–24, doi: 10.1109/ISSNIP.2014.6827618.

[38] S. Gomari, C. Knoth, and C. Antoniou, "Cluster analysis of parking behaviour: A case study in Munich," *Transp. Res. Procedia*, vol. 52, pp. 485–492, Jan. 2021, doi: 10.1016/j.trpro.2021.01.057.

[39] S. Gomari, C. Knoth, and C. Antoniou, "Spatio-temporal analysis of parked-in and parked-out events: A case study of Munich," in *Proc. Virtual ITS Eur. Congr.*, 2020, pp. 1–12.

[40] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Oct. 2011, doi: arXiv:1201.0490v4.

[41] S. Gutmann, C. Maget, M. Spangler, and K. Bogenberger, "Truck parking occupancy prediction: XGBoost-LSTM model fusion," *Front. Future Transp.*, vol. 2, pp. 1–17, Jul. 2021, doi: 10.3389/ffutr.2021.693708.

[42] S. Gomari, C. Knoth, and C. Antoniou, "Prioritization-based subsampling quality assessment methodology for mobility-related information systems," *IET Intell. Transp. Syst.*, vol. 16, no. 5, pp. 602–615, 2021. [Online]. Available: https://doi.org/10.1049/itr2.12160

[43] "Zoom levels and tile grid." Microsoft. 2020. Accessed: Jun. 20, 2019. [Online]. Available: https://docs.microsoft.com/en-us/azure/azure-maps/zoom-levels-and-tile-grid?tabs=csharp#quadkey-indic%5C