



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

Syrus Gomari^{1,2}  | Christoph Knoth³ | Constantinos Antoniou¹ 

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

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2022 The Authors. *IET Intelligent Transport Systems* published by John Wiley & Sons Ltd on behalf of The Institution of Engineering and Technology

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

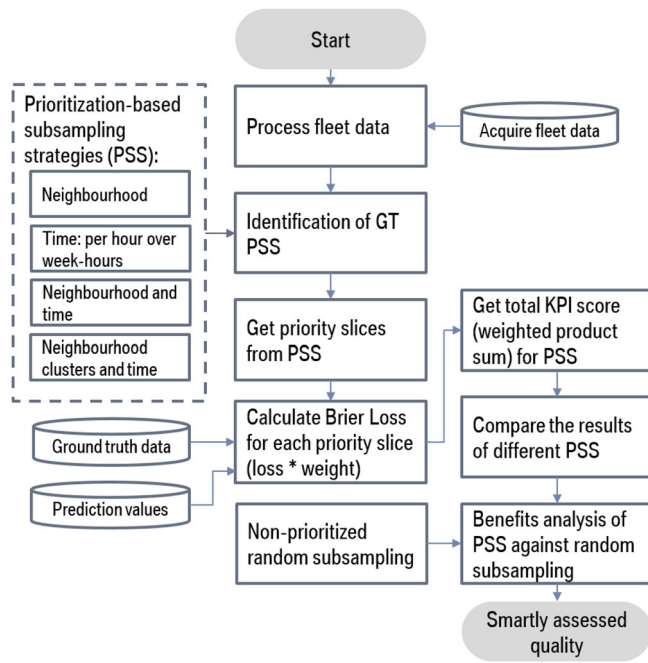


FIGURE 1 Methodology workflow

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:

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

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.

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

$$w_s = \frac{PEVolume_s}{\sum_{s=1}^N PEVolume_s}, \quad (3)$$

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

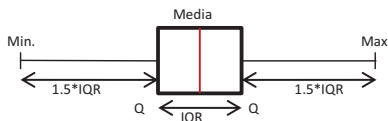
The experimental design for random subsampling of ground truth was necessary to assess and ensure the robustness of the PSS method. One objective was to ensure that if any of the PSSs are followed for ground truth collection, they can be representative of the actual quality of a prediction system. The 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.

$$IQR = Q3 - Q1, \quad (4)$$

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



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

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

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 be 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

Setup #	PSS #	Neighbourhood zoom level				TPPD cluster zoom level	Time slice
		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

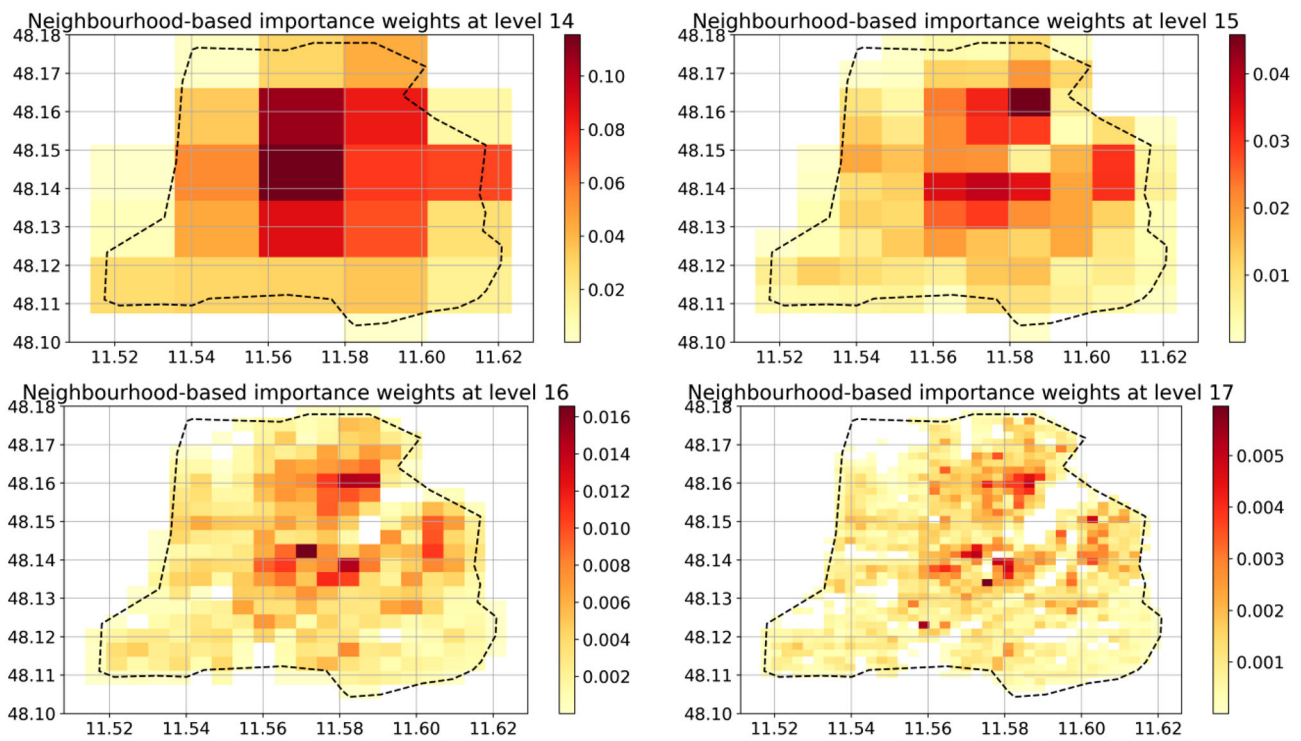


FIGURE 2 The weight importance distribution in neighbourhoods for PSS 1

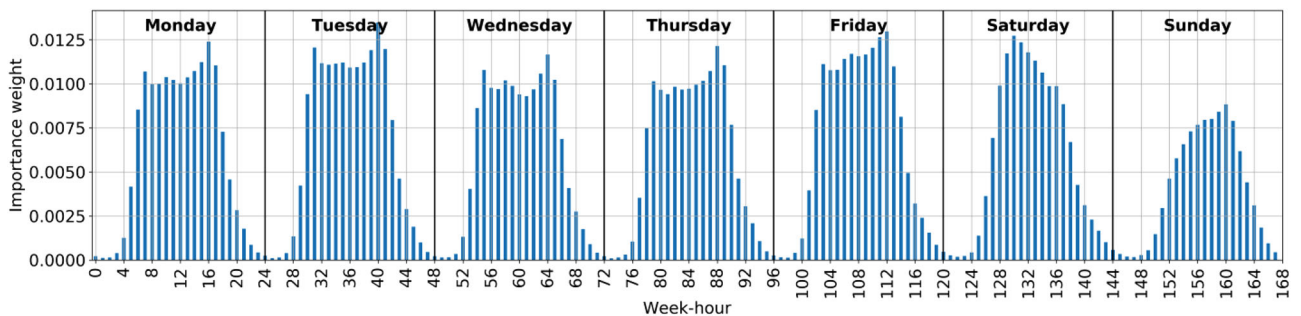


FIGURE 3 The weight distribution importance during the week-hour for PSS 2

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

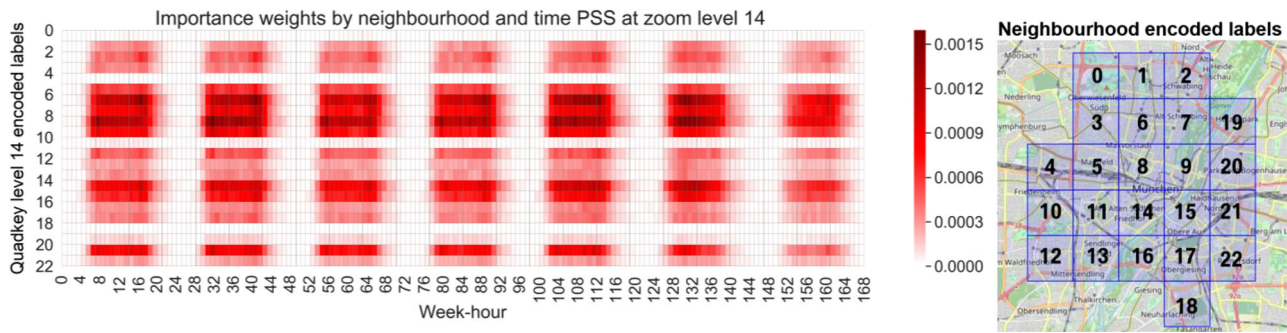


FIGURE 4 Importance weight distribution by PSS 3 on neighbourhood zoom level 14 and time (left); and encoded neighbourhood labels within Munich (right)

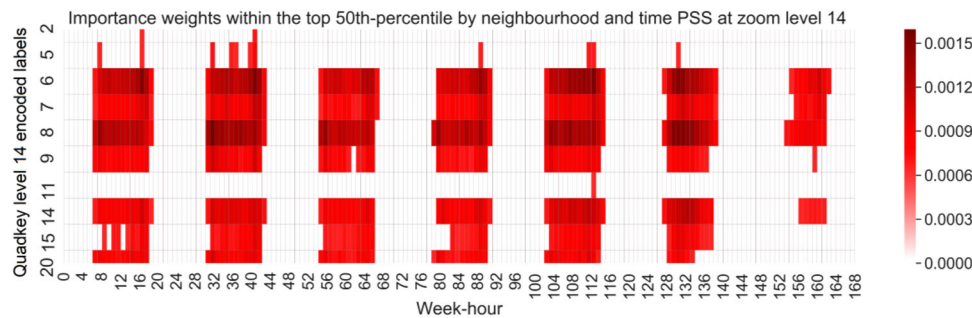


FIGURE 5 Importance weight of the same PSS 3 but for weights within top 50th-percentile

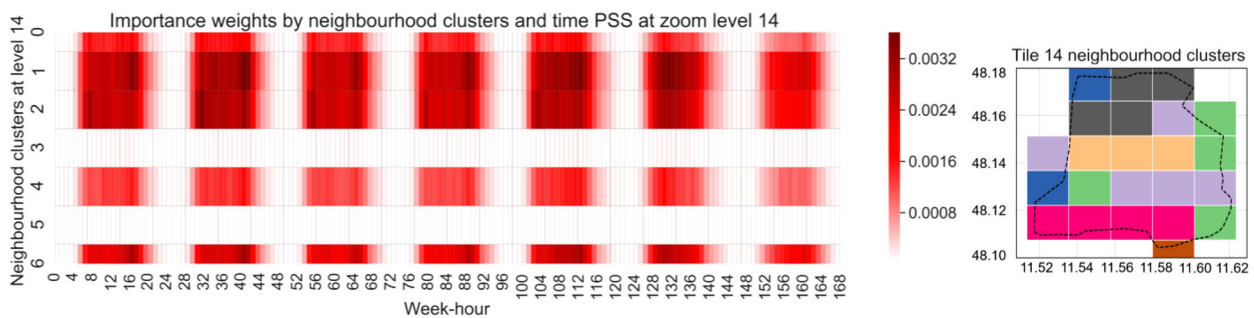


FIGURE 6 Importance weight distribution by PSS 4 on neighbourhood clusters and time (left); spatial distribution of neighbourhood clusters (right)

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

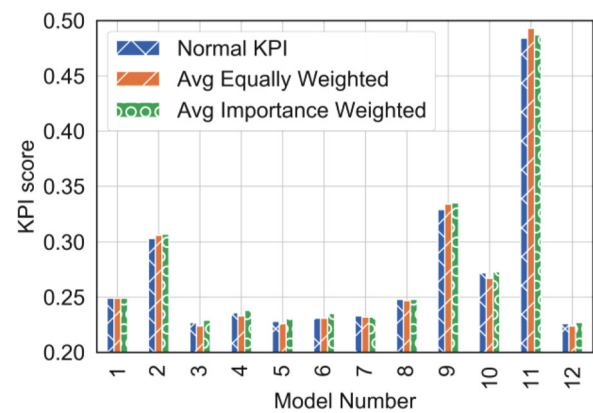
Model	Algorithm	Features	KPI	Avg. KPI	
				Eq.	Imp.
1	Xgboost	T	0.249	0.249	0.249
2	Random Forest		0.303	0.306	0.307
3	Xgboost	T, C, L	0.227	0.224	0.229
4	Random Forest		0.236	0.233	0.238
5	Xgboost	Model 3 features + hTTPD	0.228	0.226	0.231
6	Random Forest		0.231	0.231	0.235
7	Xgboost	T, h-TTPD, rt-TTPD	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 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

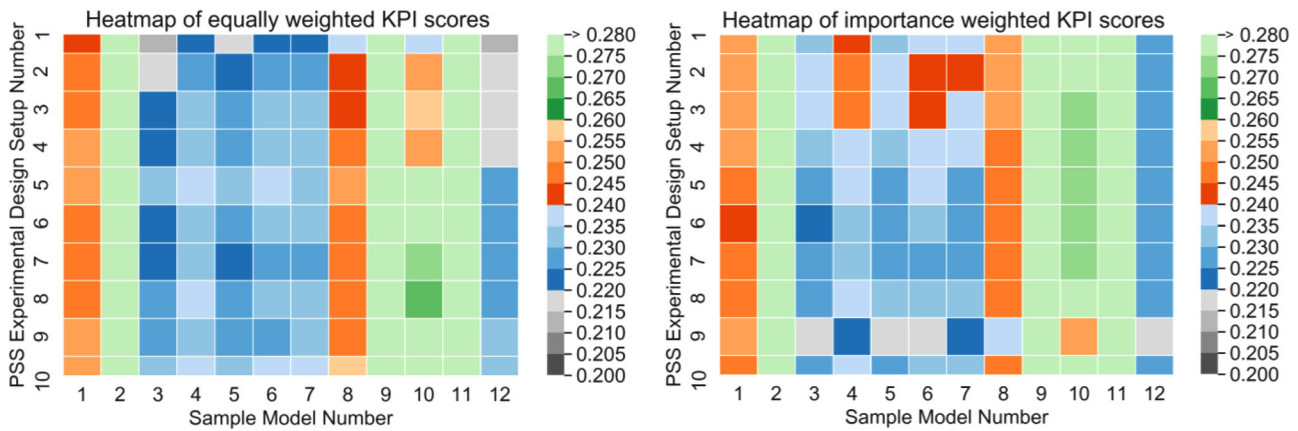


FIGURE 8 Heatmaps of equally (upper) and importance (lower) weighted KPI scores for all models considering each PSS

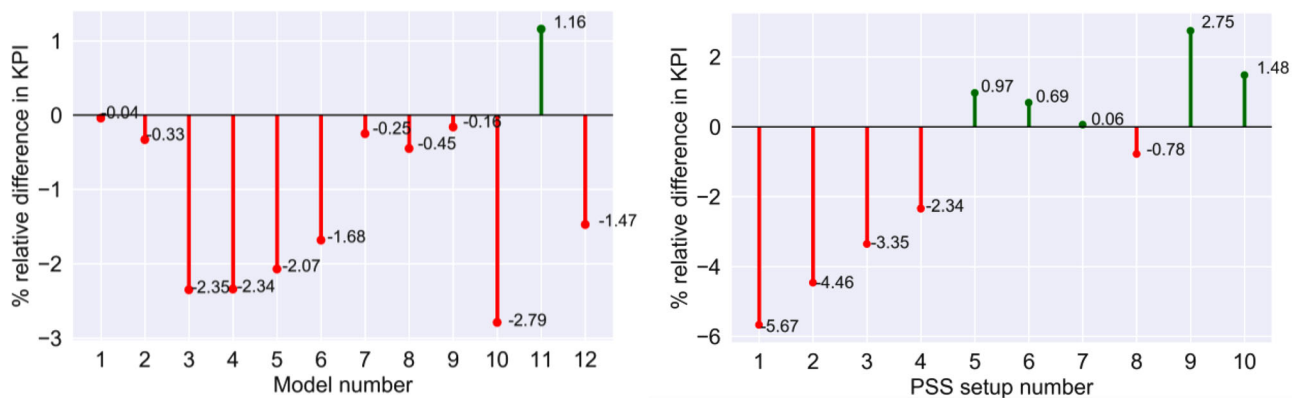


FIGURE 9 Average relative difference between equally and importance weighted KPI scores based on models (upper) or PSS setup (lower)

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		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	

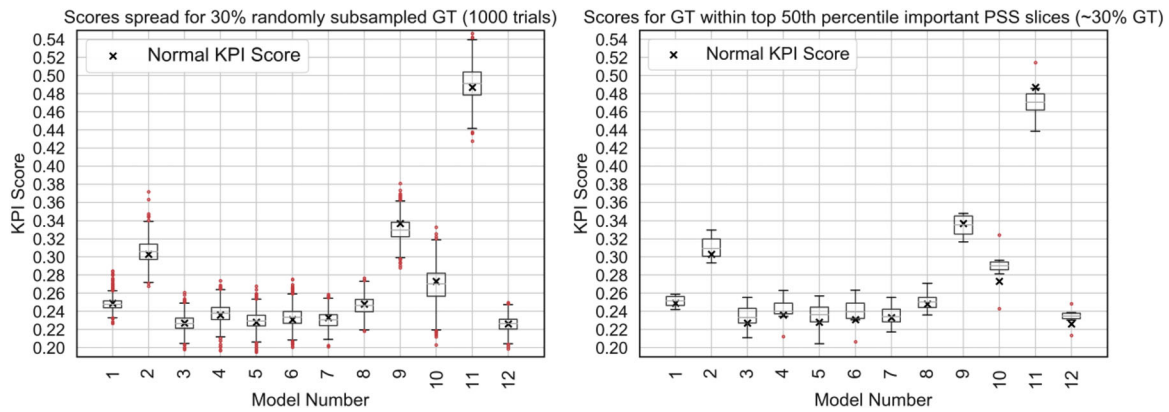


FIGURE 10 Comparison of KPI scores from NPRS against top 50th percentile important PSS

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 unfortunately 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., Sebatu, 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. Gkolas, 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>