Predicting Parking Demand with Open Data

Thomas Schuster[1] and Raphael Volz[1][0000-0003-2816-1618]

[1] University of Pforzheim, 75175 Pforzheim, Germany
raphael.volz@hs-pforzheim.de

Abstract. This paper focuses on demand forecasts for parking facilities. Our work utilizes open parking data for predictions. Several cities in Europe already publish this data continuously in the standardized DATEX II format. Traffic related information will become more ubiquitous in the future as all EU-member states must implement real-time traffic information services including parking status data since July 2017 implementing the EU directives 2010/40 and 2015/962. We demonstrate how to extract reliable and easily comprehensible forecast models for future-parking demand based on open data. These models find multiple use cases not only on a business planning level and for financial revenue forecasting but also to make traffic information systems more resilient to outages and to improve routing of drivers directing them to parking facilities with availability upon predicted arrival. Our approach takes into consideration that the data constitutes irregular time series and incorporates contextual information into the predictive models to obtain higher precision forecasts.

Keywords: : Parking prediction, machine learning, data mining, smart cities.

1 Introduction

Congestion of transport systems is a major and increasing pain point in large cities. In the vision of smart cities this issue is tackled with "computerized systems comprised of databases, tracking, and decision-making algorithms" [1] instead of brick-and-mortar extensions of infrastructure. [2] cites multiple studies claiming that up to 30% of inner-city traffic can come from drivers searching for a free parking space.

To reduce this type of traffic bigger cities typically provide stationary car-park routing systems that indicate to drivers where spaces are currently available. A more modern approach is additional online publication of this data. While some cities, such as San Francisco, already provide app-based mobile information that can travel with the driver. This approach also incorporates current availability information directly into routing. For longer distance journeys a forecast of future availability is required, which can be determined using predictive models trained on historical parking demand information. Such predictive models are also useful in case of system outages providing a second means to assess the current status.

Predictive models for demand have also a great value for parking operators and can be used to develop forecasts of revenue [3] or to improve prices through performance pricing [4]. Recommendations for variable pricing of parking are around for a long

time [5]. Understanding the demand of competitors is also beneficial to improve parking policies [6].
With improved traffic management in mind, the European Commission has required member countries to contribute to the co-ordination of traffic management and development of seamless pan European services through the ITS directive 2010/40/EU . Among many other things this directive also mandates the publishing of traffic-related open data including parking status information for important cities and parking areas along highways.
This paper focuses on demand forecasts for parking facilities based on a continuous recording of the parking status updates of 42 parking facilities in Düsseldorf between March and August 2015. Working towards a benchmark of predictive model algorithms for parking demand predictions, we show that reliable and easy to understand forecast models for future parking demand can be mined from this open data using classical statistical learning approaches. Our approach takes into consideration that the data constitutes irregular time series and incorporates contextual information into the predictive models to obtain higher precision forecasts.

The paper is organized as follows: Section 2 describes the data set and its characteristics. Section 3 details our objectives and common assumptions for predicting parking. Section 4 reports on our results using linear regression as a technique that forms a baseline for comparison. Section 5 discusses our experiment with decision trees. Section 6 summarizes our findings and provides citywide results. We conclude in Section 7 with discussion of related work and our next steps.

2 The Data Set

2.1 Data Format, Source and Content

The European Commission has sponsored the development of the XML-based DATEX format to enable traffic-related information exchange. DATEX II version 2.3 published in December 2014 now also provides a standard data model for parking information[1]. This parking information is published in two separate files, one of which provides static information (metadata) such as name, typical capacity and location of a parking facility as well as categorization into city areas. The other file provides dynamic information that contains current status, capacity and demand, absolute number of parkers, as well relative occupancy and trend information both for single parking facilities and aggregates for city areas.

While individual parking facilities could publish their status using this format, it is more common for cities to aggregate the data from several parking facilities within its city limits to provide drivers with an overview of current availabilities. For example, Tampere (Finland) publishes the current parking status since February 2015 directly

[1] http://www.datex2.eu/news/2014/12/01/datex-ii-version-23-available-now

on the Web[2]. However, the EU delegated regulation mandates every member state to maintain national access points to such data that aggregates the data from various data providers in each country.

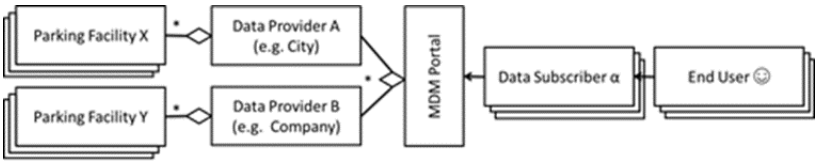


Fig. 1. A Data flow of parking information through the MDM portal (Legend: ◇ Aggregation)

In Germany this national access point is provided by the "Mobilitätsdatenmarktplatz" (MDM portal)[3] and operated by the Bundesamt für Strassenbau (BaSt) on behalf of the ministry of transport. Figure 1 shows how traffic information flows from providers to subscribers and end users through this portal. Data sets of data providers are published on a secured URL endpoint by the MDM portal. Currently 6[4] German cities are providing parking demand information for their most important parking facilities. While data access may be restricted - and the MDM portal provides mechanisms for data providers to register data subscribers and provides access control - some cities such as Düsseldorf provide open[5] data. Each data provider of the MDM portal can individually choose a frequency in which the data is updated. For example, Düsseldorf updates its dynamic parking information every minute. Each update typically only contains the changes of parking demand since the last update, and some parking facilities might not have changes in that period. Hence, subscribers must maintain memory of the parking status across individual updates to get a complete up-to-date perspective since parking facilities might provide updates at a smaller frequency than the city and may not be providing any updates for longer periods of time during closing times or because of technical difficulties. The updates received therefore constitute highly irregular multivariate time series both at the level of single parking facilities and city aggregates. Our data set is based on minutely parking data that has provided by the city of Düsseldorf for 42 parking facilities on the MDM portal in the six months between March and August 2015.

2.2 Data Processing Implementation

We receive the dynamic parking information using a scheduled job (cron job) that polls the MDM portal URL endpoint(s) on a minutely basis. The DATEX message is then parsed and its content is appended to individual CSV files to create time series

[2] http://wiki.itsfactory.fi/index.php/Tampere_Parking_DATEX2

[3] http://www.mdm-portal.de

[4] As of September 2015: Aachen (only charging stations), Düsseldorf, Frankfurt am Main, Kassel, Magdeburg, and Wuppertal

[5] Parties interested in accessing the data still need to register with the MDM portal and setup security certificates to receive data.

for each parking facility, parking area as well as one large CSV file that includes all updates received.

These CSV files are then read by data analysis scripts executed by the statistical software R [7]. In particular we use functions packages rpart [8] for decision tree fitting and zoo [9] to aggregate and plot the irregular time series that are found in the data.

2.3 Data Characteristics

In total we have received 264.376 updates in 331 MB of XML data containing 2,6 million single changes in parking status across the 42 parking facilities in our data set. 7 of these parking facilities provide incorrect data, such as negative occupancies, or only a small number of data updates, for example updates that are several days apart, and are therefore excluded from further analysis.

The number updates provided by the remaining 35 parking facilities in scope varies greatly between 21.387 and 145.104 with a median of 75.499 observations per parking facility. Hence, the average parking facility provides a status update every 3 ½ minutes to the city of Düsseldorf.

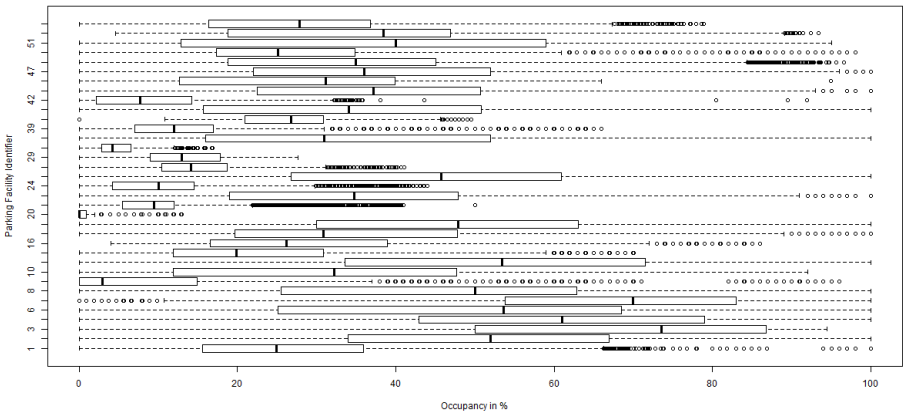


Fig. 2. Boxplot of parking facility occupancies (in %) in the data set

The box plot[6] Figure 2 shows that the variance of occupancy is very high and also that the demand characteristics vary greatly between the different parking facilities. Hence, a robust predictive model is easier to obtain per parking facility than for the city level and the city level predictive model should be computed by aggregating the predictions made for individual parking facilities.

The data of individual parking facilities shows clear periodic patterns of demand, such as illustrated by Figure 3 for two parking facilities side by side. The periodic

[6] A.k.a. box and whisker diagram showing from left to right: minimum, first quartile, median (thick line), third quartile and maximum as well as outliers (circles).

pattern of Kunsthalle Düsseldorf is similar every day (with an increased demand on Saturdays) and demand is high in late evening hours, possibly due to the fact that parking is possible for 24h and a reduced price is offered in the night. The parking facility is generally among the cheapest in the city and never really empty. Kö Galerie, a prominent luxury store in the main shopping district located at Königsallee 60, shows a clear demand pattern centered around the afternoon and almost double demand on Saturdays. Demand is neglectable when the department store is closed in the night or on Sundays.

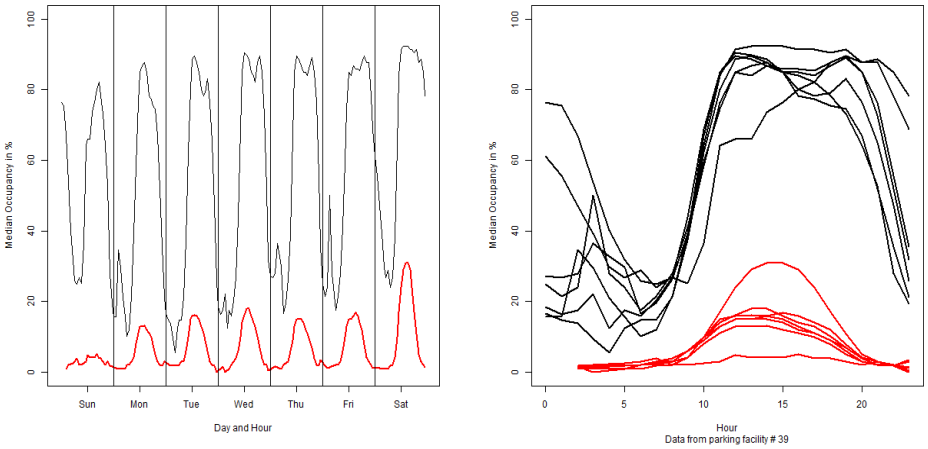


Fig. 3. Periodicity of parking demand with facility-specific temporal patterns (above #3 Kunsthalle (black) and below #39 Kö Gallerie (red))

3 Predicting Parking Demand

3.1 Objectives of our approach

Since our dataset is novel and the data has become publicly available just recently we want to establish a baseline with linear regression, a classical statistical supervised learning technique, and successively benchmark our results with other more recent machine learning and data mining approaches. In scope of this paper we will compare our linear regression baseline with decision trees, another well-known and broadly implemented regression technique.

The two main reasons for choosing those approaches is that both predictive models are very easy to comprehend for decision makers and can be implemented with few effort and computational demand either by a linear equation (linear regression) or nested if-else conditions (decision tree) in consumer apps or car information systems.

3.2 Decomposing date and time

As we have seen in Figure 3 parking demand is typically periodic but exposes complex seasonality. The periodicity of different parking facilities is highly individual due to differences in opening and closing times and different intervals between data updates. System outages that can occur at any party involved in the data chain (see Figure 1) create an additional source of irregularities in the data. Additionally, the timestamps of parking status have millisecond precision.

We therefore decompose the date and time information provided by the timestamp of each parking status update into variables for each component of time. Additionally, we determine the corresponding weekday for each date. *Hour* and *Weekday* are then used as input to the statistical learning algorithms, while all other components of time are ignored. We do not aggregate the data by hour to avoid biases.

3.3 Adding contextual information

[10] show that weather conditions influence parking demand. We therefore add hourly *rainfall* (in mm) and *temperature* (in degrees C°) data as recorded by the German national weather service (DWD) in its downtown Düsseldorf weather station. Hourly weather data is freely available[7] as part of the public mandate of DWD throughout Germany and world-wide from various data providers.

[11] discuss that public holidays influence electricity demand significantly. We therefore add binary information to the data set whether a given day is a *public holiday*, *school holiday* or a *Brueckentag*[8]. 35% of the observations in our data set are recorded on either one of these special days.

Both contextual data sets are available ahead of making predictions as 24h weather forecasts and averaged historical weather observations as well as holiday calendars that are planned many years ahead (school holidays) or regulated by legislature (public holidays). We will compare how our predictive models perform with and without this contextual data.

3.4 Identifying the dependent variable

Parking facilities provide information on their status by transmitting both relative *occupancy* and absolute *occupied* spaces as well as *vacant* (available) spaces and their current *capacity* that fluctuates, for example due to ongoing renovations or by systematically opening and closing levels based on demand. Obviously, this information is redundant[9] and highly correlated.

---

[7] http://www.dwd.de/WESTE (free registration of user account required)

[8] A *Brueckentag* is a single working day that fall between a public holiday and a weekend where many Germans take a day off from work

[9] Occupancy = occupied / capacity and vacant = capacity – occupied.

Since the capacities of the parking facilities vary greatly, we choose to predict the absolute number of *occupied* spaces. *capacity* and the number of *long-term* parkers, that some garages report, are also independent variables in the predictive models.

3.5 Splitting Training and Test Data

We subset all 2.6 million observations into separate smaller data sets that single out observations per parking facility. Splitting the data sets into training and test data sets by months as often found in related work will lead to biases due to the different number of weekdays, length of months and occurrence of holidays. We therefore split the facility-specific data sets into 75% training and 25% test data in an algorithmic fashion[10] that preserves the relative ratios of the dependent variable in both data sets. The same training and test data sets are used to benchmark our approaches.

3.6 Assessing predictive model quality and comparing models

We assess the coefficient of determination R^2 for each predictive model on the training and test data. We indicate how well the models fit to the training data, and observe whether this fit deteriorates when making predictions on the unknown test data (indicated by a lower R^2).

To assess the overall quality of the model and compare model quality between experiments as well as between parking facilities we define the relative root mean-square deviation ($rRMSD$), which normalizes the well known $RMSD$ by capacity. The $rRMSD_i$ is calculated for each parking facility i based on the predicted values $\hat{y}_t$ for times t of the dependent variable y (occupied) for all n different predictions as the square root of the mean of the squares of the deviations and normalized by the capacity of the parking facility k_i (1).

$$rRMSD_i = \frac{\sqrt{\sum_{t=1}^{n} \frac{(\hat{y}_t - y_t)^2}{n}}}{k_i} \quad (1)$$

The rRMSD allows comparing parking facilities with different capacities and denotes the relative prediction error of a predictive model (in % of capacity). rRMSD is an unbiased estimator like RMSD.

4 Linear Regression Models

We build similar linear regression models with ($\hat{y}_+$) and without ($\hat{y}_-$) contextual data for every parking facility. The regression model without context predicts the number of occupied parking spaces $\hat{y}_-$ based on the independent variables occurring in the linear equation v, where factor variables hour $\vec{h}$, weekday $\vec{w}$, and numeric variables

[10] Using sample.split function offered by caTools (Tuszynski 2014).

of current capacity k and number of long-term parkers l are used as inputs, adjusted by the intercept c_-.

$$\begin{aligned} \hat{y}_- &= v + c_- \\ v &= \vec{c}_h \vec{h} + \vec{c}_w \vec{w} + c_k k + c_l l \end{aligned} \tag{2}$$

The model with context predicts occupied parking spaces $\hat{y}_+$ in a similar fashion and adds numeric variables for temperature t, rainfall r as well as binary variables for Brueckentage b, school holidays s, and public holidays f, adjusted by the intercept c_+.

$$\begin{aligned} \hat{y}_+ &= v + a + c_+ \\ a &= c_t t + c_r r + c_b b + c_s s + c_f f \end{aligned} \tag{3}$$

All coefficients c_i in equations (2) and (3) are determined by the supervised learning task based on the training data, where factorial coefficients $\vec{c}_j$will take different values for every factor. The coefficients will be different for every predictive model and adjust to the specific parking facility data.

For example, Table 1 shows the coefficients learned for parking facility Kö-Gallerie (#39) to occupied spaces $\hat{y}_+$ with the contextual information. The basic estimate c_+ of occupied parking spaces at 2 pm is 121. Thus, we expect 121 more cars to be present, while at 11 pm only a total of $121 - 28 = 93$ cars are expected.

Table 1. Excerpt of the linear regression model with context predicting occupied spaces $\hat{y}_+$ for parking facility Kö-Gallerie (#39)

(Factorial) coefficient		Estimate	Std. Error	t value
c_+		121.06	4.10	29.55
$\vec{c}_h$	hour3	16.92	24.91	0.68
	…			
	hour14	121.12	3.10	39.12
	…			
	hour23	-28.13	8.79	-3.20
$\vec{c}_w$	weekday1	56.45	0.90	62.71
	…			
	weekday6	134.28	0.88	152.03
c_k		-0.21	0.00	-60.13
c_t		0.49	0.03	15.81
c_r		0.63	0.39	1.62
c_f		-63.42	1.77	-35.87
c_s		-10.17	0.36	-28.40
c_b		120.63	2.21	54.70

Similarly, on Saturdays[11] we generally expect 134 more cars. Both higher temperatures c_t and rainfall c_r increase occupancy by one car for every 2°C and 3 cars for every 2 mm of rain. 63 cars less can be expected on a public holiday and 10 cars less on school vacations while 120 more cars are parking on Brueckentage.The predictive model for Kö-Gallerie is robust and has a R^2 of 0.68 (without context) and 0.71 (with context) on the training data and 0.69 / 0.71 on the test data with a $rRMSD$ of 6.9% (without context) and 6.4% (context). Adding context provides a 4% improvement in prediction.

Table 2 reveals that R^2 the predictive models are generally robust on the test data. By comparing contextual and context-free predictions and excluding both best and worst models, it displays large deviations between 22% and 6.2% prediction error ($rRMSD$) for models without context that only slightly improve when context is added to 21% and 5.6%. In particular, context does not make bad models much better nor good models any worse. Overall adding contextual data provides 5.2% improvement in predictive quality on the capacity-weighted average of all linear regression models.

Table 2. Overall quality of linear regression models across 35 parking facilities

	$\hat{y}_-$ (without context)			$\hat{y}_+$ (with context)		
	Training R^2	Test R^2	$rRMSD$	Training R^2	Test R^2	$rRMSD$
2nd Worst $\hat{y}_-/\hat{y}_+$	0.06	0.06	22%	0.17	0.17	21%
Median $\hat{y}_-$	0.54	0.54	13.4%	0.56	0.57	13.1%
Median $\hat{y}_+$	0.56	0.56	13.9%	0.57	0.57	13.7%
2nd Best $\hat{y}_-/\hat{y}_+$	0.85	0.85	6.2%	0.88	0.88	5.6%

5 Decision Trees

The decision tree models [12] are trained with the same data set and input as the linear regression models, in particular we train two variants without (2) and with (3) contextual information. Whether input variables are used in the model depends on the results of the tree-fitting algorithm for the particular parking facility. Across all parking facilities, we obtain fits with the same parameters. To avoid overfitting to the data a minimum number of 30 observations must be in any terminal leaf node of the decision tree. We additionally avoid additional tree splits when the overall R^2 does not increase by at least $\frac{1}{1000}$.

The decision trees fitted from the training data for each parking facility are easy to understand but cumbersome to read and best-fitted for automated decision support. This is due to the fact that an average of 67 (maximum of 120) decision criteria are involved for trees trained on data with context. Figure 4 therefore only shows a sample tree that demonstrates important characteristics that are shared among the actual trees fitted to the training data. Top-level distinctions are typically made based on

[11] Saturday is day 6 counting from 0 as Sunday, etc.

hour $\vec{c}_h$ and weekday $\vec{c}_w$ variables. Without considering any input variables, evaluating the tree would predict 107 occupied parking spaces. The first binary distinction is made based on whether the hour $\vec{c}_h$ is either before 10 am after 6 pm. If so, we should assume only 61 occupied spaces. If not, we can assume 137 occupied spaces.

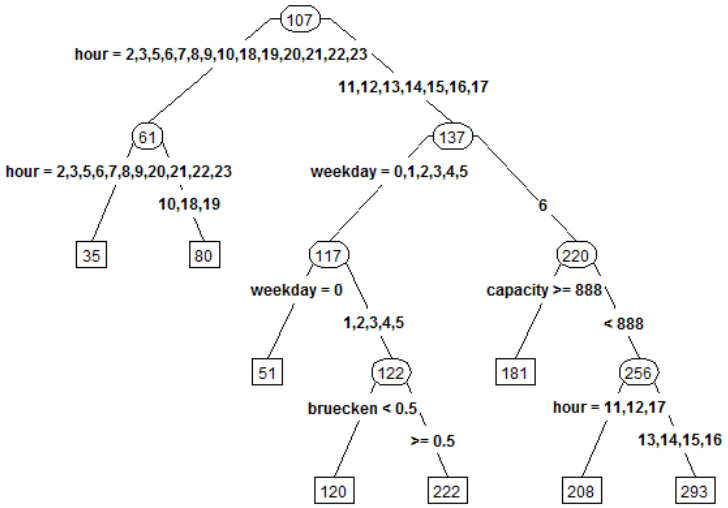


Fig. 4. Sample binary decision tree for Kö Gallerie parking facility (#39)

These decisions are recursively refined while walking towards the leaves of the tree turning left when a condition is met and right if not until a leaf node is reached. For example, we assume 293 occupants between 1 pm and 4 pm on Saturdays, if our capacity is below 888 parking spaces.

Table 3 shows that the R^2 of the decision trees are also generally robust on the test data. We again exclude the best and worst models and observe smaller deviations between 14.6% and 5.7% prediction error ($rRMSD$) for models without context. Adding context generally improves the models and can observe between 12.8% and 4.3% prediction error. Adding contextual data provides a 15.6% improvement in predictive quality on the capacity-weighted average of all decision tree models.

Table 3. Overall quality of decision tree models across 35 parking facilities

	$\hat{y}_-$ (without context)			$\hat{y}_+$ (with context)		
	Training R^2	Test R^2	$rRMSD$	Training R^2	Test R^2	$rRMSD$
2nd Worst $\hat{y}_-$/ $\hat{y}_+$	0.53	0.53	14.6%	0.70	0.69	12.8%
Median $\hat{y}_-$	0.8	0.8	10.4%	0.87	0.87	8.3%
Median $\hat{y}_+$	0.42	0.41	11.2%	0.58	0.57	9.64%
2nd Best $\hat{y}_-$/ $\hat{y}_+$	0.79	0.79	5.7%	0.88	0.88	4.3%

6 City-wide predictions and overall results

We define the citywide prediction error $rRMSD_{city}$ as a capacity-adjusted aggregate that normalizes the aggregated $rRMSD_i$ by the total capacity k_i across all n parking facilities:

$$rRMSD_{city} = \frac{\sum_{i=1}^{n} k_i \cdot rRMSD_i}{\sum_{i=1}^{n} k_i} \quad (4)$$

Table 4 compares the overall prediction error $rRMSD_{city}$ across all 21.381 parking spaces in the data set. We can observe that contextual data generally improves prediction models in both approaches and provide a greater improvement to decision trees. Likewise, we can observe that decision tree models outperform linear regression models by 27.6% on the citywide level.

Table 4. Citywide prediction error $rRMSD_{city}$

	Linear Regression	Decision Tree	*Improvement*
$\hat{y}_-$ (without context)	**8.9%**	**7.2%**	*18.7%*
$\hat{y}_+$ (with context)	**8.4%**	**6.1%**	*27.6%*
Improvement	*5.2%*	*15.6%*	*31.4%*

We can generally confirm these improvements for every single predictive model although the improvements vary significantly across the individual parking facilities. Since $rRMSD_{city}$ is smaller than the median prediction error we can conclude that our predictive models generally perform better for large parking facilities.

7 Conclusion

We have shown that robust predictive models for demand of parking facilities can be obtained from open parking status data and that classic regression techniques readily generate predictions with acceptable error rates while providing easy to understand models for decision makers and easy to implement equations and rules to implement decision-making algorithms. We have seen that contextual data related to (public and school) holidays and weather data can decrease prediction errors significantly and also show higher coefficients of determination R^2.

We are currently pursuing several directions to expand our results. First, we intend assessing the stability of the models with more data. This does include (a) stabilization with data of other cities with published DATEX II parking status and (b) addressing the challenges of irregularities by compensating failures in data provision with similar data providers.

We are currently benchmarking our results with other algorithms for multivariate time series with complex seasonality from the statistical community such as multivar-

iate volatility models [13] as well as other supervised algorithms from the machine learning community that have been proposed for parking predictions such as wavelet neural networks [14].

Another direction for future research will be to develop an approach for short-term demand forecasts that considers the last known status as well as expected progressions from this status into the next couple of hours.

We will publish the data set after the anonymous review of this paper in an open data repository to provide a benchmark for parking prediction models since prior work on parking demand forecasts is based on proprietary data sets and our data set is the largest generally available dataset encompassing data from several parking facilities.

Acknowledgements: We thank the city of Düsseldorf to provide the data set openly and an anonymous company for providing technical assistance with the DATEX II format and providing us with access to the MDM portal.

8 References

1. Hall, R.E., Bowerman, B., Braverman, J., Taylor, J., Todosow, H., Von Wimmersperg, U.: The vision of a smart city. Brookhaven National Lab., Upton, NY (US) (2000).
2. Shoup, D.: The High Cost of Free Parking, Updated Edition. Routledge, Chicago (2011).
3. Burns, M.R., Faurot, D.J.: An econometric forecasting model of revenues from urban parking facilities. Journal of Economics and Business. 44, 143–150 (1992).
4. Pierce, G., Shoup, D.: Getting the Prices Right. Journal of the American Planning Association. 79, 67–81 (2013). https://doi.org/10.1080/01944363.2013.787307.
5. Vickrey, W.: The economizing of curb parking space. Traffic Engineering Magazine, 62–67 (1954).
6. Arnott, R.: Spatial competition between parking garages and downtown parking policy. Transport Policy. 13, 458–469 (2006).
7. R Core Team: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (2018).
8. Therneau, T.M., Atkinson, E.J.: An introduction to recursive partitioning using the rpart routines. Divsion of Biostatistics 61. (1997).
9. Zeileis, A., Grothendieck, G.: zoo: S3 Infrastructure for Regular and Irregular Time Series. Journal of Statistical Software. 14, 1–27 (2005). https://doi.org/10.18637/jss.v014.i06.
10. David, A., Overkamp, K., Scheuerer, W.: Event-oriented forecast of the occupancy rate of parking spaces as part of a parking information service. In: Proceedings of the 7th World Congress on Intelligent Systems (2000).
11. Livera, A.M.D., Hyndman, R.J., Snyder, R.D.: Forecasting Time Series With Complex Seasonal Patterns Using Exponential Smoothing. Journal of the American Statistical Association. 106, 1513–1527 (2011). https://doi.org/10.1198/jasa.2011.tm09771.
12. Breiman, L.: Classification and regression trees. Routledge (2017).
13. Tsay, R.S.: Multivariate Time Series Analysis: With R and Financial Applications. Wiley, Hoboken, New Jersey (2014).
14. Ji, Y., Tang, D., Blythe, P., Guo, W., Wang, W.: Short-term forecasting of available parking space using wavelet neural network model. IET Intelligent Transport Systems. 9, 202–209 (2014).