Statistical Analysis and Prediction of Parking Behavior

Ningxuan Feng[1], Feng Zhang[1(✉)], Jiazao Lin[2], Jidong Zhai[3], and Xiaoyong Du[1]

[1] Key Laboratory of Data Engineering and Knowledge Engineering (MOE), and School of Information, Renmin University of China, Beijing 100872, China
fengzhang@ruc.edu.cn

[2] Department of Information Management, Peking University, Beijing 100871, China

[3] Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China

Abstract. In China, more and more families own cars, and parking is also undergoing a revolution from manual to automatic charging. In the process of parking revolution, understanding parking behavior and making an effective prediction is important for parking companies and municipal policymakers.

We obtain real parking data from a big parking company for parking behavior analysis and prediction. The dataset comes from a shopping mall in Ningbo, Zhejiang, and it consists of 136,973 records in 396 days. Specifically, we mainly explore the impact of weather factors on parking behavior. We study several models, and find that the random forest model can make the most accurate parking behavior prediction. Experiments show that the random forest model can reach 89% accuracy.

Keywords: Prediction model · Regression · Weather condition

1 Introduction

Currently, China has more than 217 million cars, and has a huge demand for parking lots [12]. It becomes very important to improve the utilization of parking space because the cars have faster growth. It also increases the demand for developing intelligent parking system, which can provide better parking management and higher profits for the owners of parking lots.

In the past few years, several parking-related types of research have been conducted to improve parking from different perspectives. For example, some studies [4,23,27,28,31] aim to provide parking information to drivers for free parking; Fang and others [7] proposed an algorithm to allocate cars to parking grid, aiming to improve the utilization of parking space.

The requirement of parking space is an important part of intelligent parking; the studies above considered the prediction of the requirement. However, few

Published by Springer Nature Switzerland AG 2019
X. Tang et al. (Eds.): NPC 2019, LNCS 11783, pp. 93–104, 2019.
https://doi.org/10.1007/978-3-030-30709-7_8

of them involve weather conditions in parking prediction. In daily life, weather condition has a remarkable impact on our travel plan.

In this paper, we analyze the parking behaviors with weather considered, and then explore various models for parking prediction. In detail, we obtain real parking dataset from a big parking company for parking behavior analysis and prediction. The dataset comes from a shopping mall in Ningbo, Zhejiang, and it consists of 136,973 records in 396 days. We consider the influence of temperature, humidity, rainfall and wind speed. We use the Anova test [9] to analyze different categorical features, and test the correlation between all numerical features by pair plot. Moreover, we also separate workdays from holidays.

For the parking behavior prediction, we have explored linear regression [26], ridge regression [14], Lasso regression [10], decision tree [24], and random forest [15] to depict parking behaviors with weather considered. We find that the random forest is the most suitable model for parking behavior analysis and prediction. Experiments show that it achieves 94% accuracy; its root mean square error (RMSE) can be narrowed down to 0.1662, which is smaller than the other models.

2 Background

2.1 Parking Behavior

Parking behavior refers to the range of actions and mannerisms related to parking. In this paper, we mainly refer to the number of parking each day. In our life, traveling out with cars and demand for off-car activities lead to parking behavior. The purpose of parking can be business, shopping or accommodation. Parking behavior has increased significantly in recent years because of the rapid growth of the number of cars.

The parking behavior is changeable because it can be affected by many factors, especially weather. When it rains heavily, people would more likely to choose traveling out with cars if the activity is necessary. There are also other important determinants for travel plan related to parking behavior. For example, in holidays, the location of the parking lot also has a great influence on parking behavior.

2.2 Motivation

Prediction is meaningful in many fields, not only in computer architecture [16,32], but also in parking behavior [18,21,29]. Parking behavior plays an important role in the city's traffic management. Policymakers can optimize the traffic control strategy in real time based on parking behavior, such as changing the duration of some traffic lights. For parking lot managers, accurate prediction of parking behavior helps develop policies that can improve parking space utilization and get more benefits.

Predicting parking behavior makes a lot of sense. Several related works have been developed in recent years [1,18,21,29]. These studies proposed models to

predict parking space availability and occupancy, which partially depicts parking behavior. However, none of them consider the influence of weather condition on parking behavior. This paper is the first to involve weather in parking behavior analysis and prediction.

2.3 Challenges

To conduct an extensive study of parking behavior, we face three major challenges.

Challenge 1: Irregular Data. The data we used to train the prediction model is disorganized. To eliminate the effect of impurity, we need to fully understand the data, and conduct specified data cleaning.

Challenge 2: Various Weather Factors. Weather condition is composed of many detailed factors, such as temperature, humidity, and wind speed. They all affect the prediction accuracy of parking space demand.

Challenge 3: Model Selection. Since there is no research before for weather-related parking behavior analysis and prediction, it is difficult to select the most appropriate model for training.

3 Solution Overview

3.1 Experimental Setup

In this paper, our parking dataset is composed of the parking records of 21-Wharf shopping mall parking lot in Ningbo City, Zhejiang Province, China. The dataset spans 13 months from March 1st, 2018 to March 31th, 2019. It consists of 136,973 parking records. For each parking records, we obtain parking information including the starting time, the ending time, and the selected parking space identity number. We regard the parking record with a duration less than five minutes as noise data. The weather dataset is weather-by-hour data for Ningbo. For each hour, we got precipitation, temperature, relative humidity, and wind speed. We also add some extra categorical features into the dataset that may potentially influence the analysis and prediction. These features include holiday, month, year, weathersit (decided by precipitation), weekday, and season.

3.2 Analysis and Prediction Framework of Parking Behavior

As stated in Sect. 2.3, we have three major challenges, irregular data, various weather factors, and model selection. For the first challenge, we visualize the data to assess the distribution of features, and then present a regularization to reduce the effect of impurity. For the second challenge, we implement a feature selection module to find out whether all the features are necessary for training, and eliminate the outliers. As to the last challenge, we explore five models for park behavior prediction.

The analysis and prediction framework consists of three modules, (1) data preprocessing module, used for data visualization and regularization (Sect. 4), (2) feature selection module, used to clean data and select major features (Sect. 5), and (3) parking space modeling, which explores related models.

4 Preprocessing Methodology

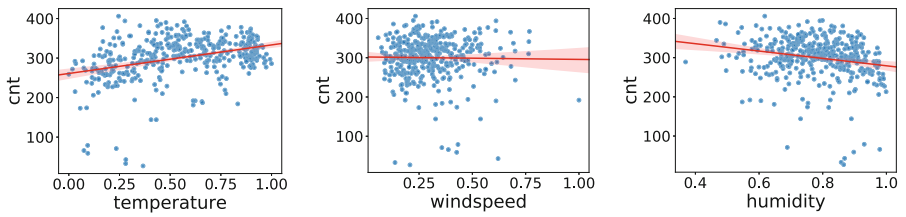
In order to perform an overall analysis of the relevance, we first perform visualization for both parking records and the related features. The features can be divided into two categories: numerical features and categorical features. The numerical features include temperature, wind speed, and humidity, which can be represented as numbers. The categorical features are features that belong to some categories, such as season, working day or holiday, and weather categories (sunny, windy, rainy, and so on).

4.1 Numerical Features

In this part, we analyze the numerical features and use temperature, wind speed, and humidity for illustration. We first normalize features using Eq. 1, and then check for Gaussian distribution [22]. According to our observation, the distribution of these features is in accordance with Gaussian distribution.

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{1}$$

We show the scatter plot of numerical features versus car parking count (denoted as `cnt`) in Fig. 1. Figure 1(a) exhibits the relation between `normalized temperature` and `cnt`. It shows that as the temperature increases, the `cnt` also increases, and the relation between `temperature` and `cnt` has a positive relationship, though there are some outliers.



(a) Temperature vs cnt. (b) Windspeed vs cnt. (c) Humidity vs cnt.

Fig. 1. Linear regression model fit of numerical features to `cnt`. The line represents the regression trend.

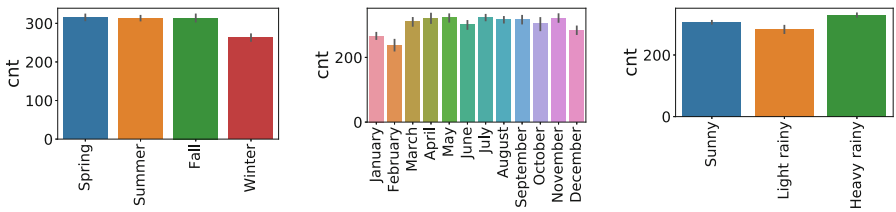
Figure 1(b) shows a scatter plot of `normalized wind speed` versus `cnt`. We can see that when we compare the feature alone with `cnt`, the distribution is

little scattered with concentration mainly on the lower side of the `normalized wind speed`.

The scatter plot of `humidity` versus `cnt` in Fig. 1(c) shows that as humidity increases, `cnt` decreases, which implies that people tend to avoid parking cars in 21-Wharf shopping mall when the humidity is high.

4.2 Categorical Features

In this part, we explore categorical features, including season, year, month, holiday, and weathersit. We show the relation of categorical features versus `cnt` in Fig. 2.



(a) Season vs cnt. (b) Month vs cnt. (c) Weathersit vs cnt.

Fig. 2. The relation between categorical features and `cnt`.

For the feature of the season, it has four categories: spring, summer, fall, and winter. Our dataset includes both March 2018 and March 2019, so we have about 120 days of spring, and 90 days for the other seasons. The season-related variation of car parking in Fig. 2(a) reveals that `cnt` in winter is much less than that in the other seasons. This phenomenon infers that people may not willing to travel out in winter.

The feature `year` has two values, 2018 and 2019. Our dataset has more days from 2018 than from 2019, because there are nine months in 2018 and four months in 2019 in our dataset. However, we find that the year 2019 has more car parking on average than the year 2018 does, which probably relates to the call of low-carbon traveling.

As to the feature of `month`, Fig. 2(b) shows that some months have fewer car parking, such as January, February, and December. It indicates that people tend not to drive out in these months, which is consistent with the phenomenon of `season`.

The number of `holidays` is less than that of working days in our dataset. We count the average of the parking times, `cnt`, for holidays and working days. Our analysis shows low `cnt` for working days than for holidays, which indicates that people travel out with cars more on holidays considering this parking lot.

For the influence from the categorical feature of weathersit, we consider three categories: sunny, light rainy, and heavy rainy, as shown in Fig. 2(c). Our dataset has more sunny days than rainy days. However, we count the average of `cnt`, and it shows that `cnt` are higher in heavy rain than in the others.

5 Feature Selection

In order to choose the right set of predictors, we need to perform feature selection before applying predictors to our model. Although more features imply more information on our dataset, they also lead to higher variance. In this section, we start with the outlier analysis.

5.1 Outlier Analysis

Outliers are the data points that differ greatly from other observations, which should be removed from our dataset. In our study, we use the method in [2] to delete those data. Specially, the data points with less than 1.5 interquartile range times the 25th percentile, or more than 1.5 interquartile range times the 75th percentile, are treated as outliers. We visualise the numerical features with (such as `cnt`) and without (such as fixed `cnt`) outliers in Fig. 3.

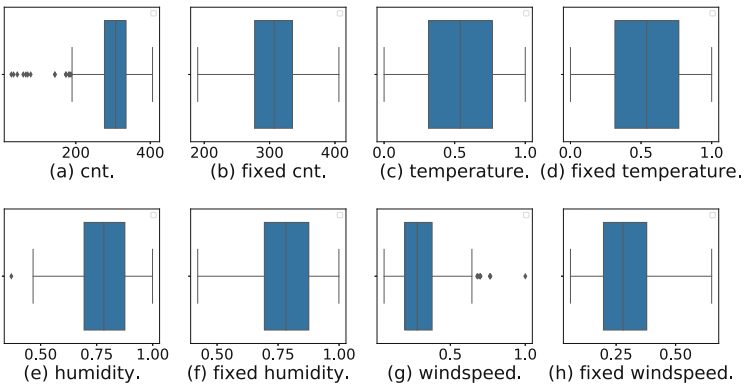


Fig. 3. Numerical features with and without outliers.

In addition, please note that the location of the parking lot also plays an important role in parking behavior. Because we only analyze one parking lot, we do not consider the location influence. We leave it to our future work.

5.2 Feature Analysis

We first show the pair plot for all numerical features in Fig. 4 to see the correlation between a pair of variables. Figure 4 shows that each pair of variables is uniformly distributed, no evident linear correlation between any pair of variables. In a word, each numerical feature is independent of the others.

As the target variable `cnt` is continuous (we turn it to continuous in the normalization of Sect. 4.1), we perform Anova (analysis of variance) [9] validation for checking the variation in the target variable explained by the categorical

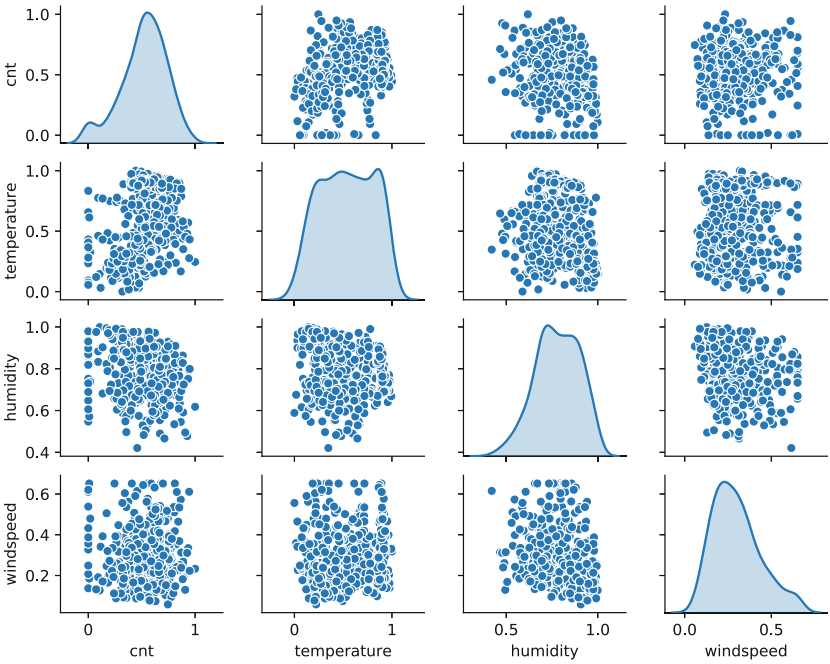


Fig. 4. Pair polt for all numerical features.

feature set. Considering 95% confidence interval, feature variables with p-value more than 0.05 shall be discarded.

We demonstrate the Anova for all categorical features in Table 1. The F-statistic represents the variation between sample means divided by the variation within the samples. It is the probability of the observed result the same as the one obtained in the experiment, assuming the null hypothesis [9] is true. Low P-values are indications of strong evidence against the null hypothesis. It can be seen from Table 1 that no feature has P-value more than 0.05.

Table 1. Anova results on categorical dataset.

Categorical feature	Season	Year	Month	Holiday	Weekday	Weathersit
F-statistic	211.46	893.92	1089.08	87.31	608.05	304.70
P-value	1.26e−42	5.70e−132	8.42e−151	9.26e−20	5.18e−100	5.90e−58

After the introduction of data preprocessing and feature selection, we have normalized the numerical features, eliminated the effects of the outlier and selected a workable set for our training. Next, we shall explore the parking behavior prediction with various models.

6 Parking Behavior Prediction

In this section, we are exploring models that can predict `cnt` with those numerical and categorical features.

6.1 Modeling Methods

Regression is widely used for prediction. In this part, we explore the following models to demonstrate their efficacy in parking behavior prediction.

- **Linear Regression Model** [26]. Given a set $\{y_i, x_{i1}, \ldots, x_{ip}\}_{i=1}^{n}$, a linear regression model assumes that the relationship between the dependent variable y and the p-vector of regressors x is linear. We also consider two generalized linear regression models: ridge regression [14] and Lasso regression [10].
- **Decision Tree** [24]. A decision support tool that uses a tree-like model of decisions and their possible consequences.
- **Random Forest** [15]. An ensemble learning method for classification, regression, and other tasks that are operated by constructing multitudes of decision trees.

6.2 Model Evaluation

In this part, we use linear regression, decision tree, and random forest models for parking behavior prediction, and use Eq. 2 to verify the model accuracy. The dataset covers 396 days. We randomly select 75% days (297 days) as training data, and 25% days (99 days) for validation.

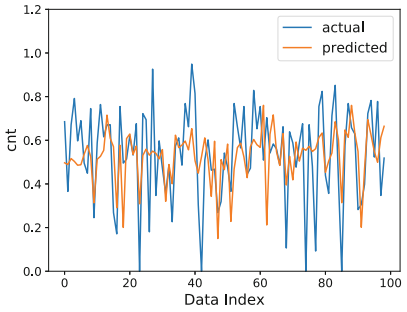
$$accuracy = \frac{|cnt_{real} - cnt_{predicted}|}{cnt_{real}} \tag{2}$$

Linear Regression Model. We first perform an Ordinary Least Squares regression (OLS) model [25] shown in Table 2. The three features with the highest absolute value of coefficient are temperature, humidity, and wind speed. Their coefficients are positive, which means that when these three features are high, the parking lot has a higher utilization. In addition, the coefficient of temperature is 0.148, which is less than the coefficient of wind speed (0.166); this shows that wind speed has a higher impact on `cnt` than temperature does.

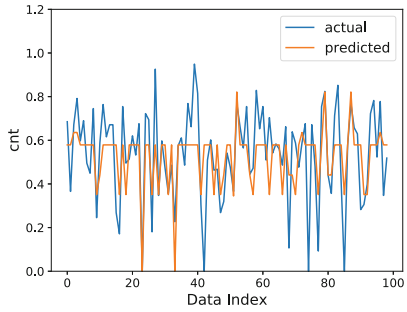
We show the output of the predictor using linear regression in Fig. 5(a). The accuracy of linear regression is 78%. In addition, ridge regression model [14] and Lasso regression model [10] are used to regularize the linear regression. We calculate the R-square and RMSE (Root Mean Squared Error) to test the predictors. For the ridge regression model, the best alpha is 0.1, the R-square is 0.3672, and the RMSE is 0.1714. We acquire similar results for the Lasso regression model with best alpha 0.001, R-square 0.3656, and RMSE 0.1719.

Table 2. OLS regression results.

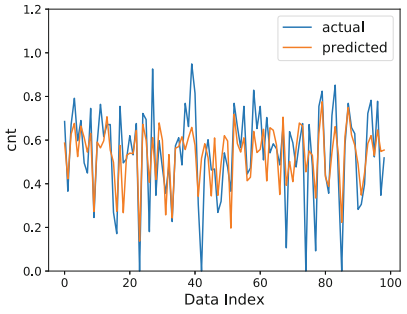
Feature	coef	std err	t	P $>\|t\|$	[0.025	0.975]
Season	−0.0694	0.014	−5.121	0.000	−0.096	−0.043
Year	−0.0090	0.033	−0.268	0.789	−0.075	0.057
Month	0.0205	0.005	4.328	0.000	0.011	0.030
Holiday	0.0892	0.024	3.768	0.000	0.043	0.136
Weekday	0.0238	0.005	4.429	0.000	0.013	0.034
Weathersit	0.0045	0.031	0.145	0.885	−0.057	0.066
Temperature	0.1483	0.050	0.2962	0.003	0.050	0.0247
Humidity	0.3534	0.079	4.496	0.000	0.199	0.508
Wind speed	0.1664	0.077	2.167	0.031	0.015	0.318



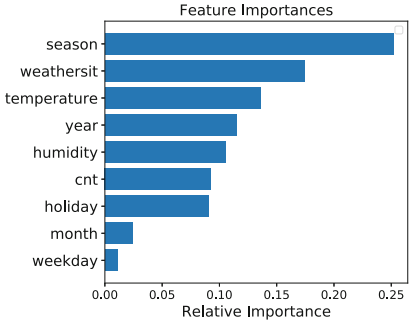
(a) Predicted result from linear model.



(b) Predicted result from decision tree model.



(c) Predicted result from random forest model.



(d) Feature ranking from random forest model.

Fig. 5. Predicted results. Data index refers to the index of records in the test set.

Decision Tree. We also use a decision tree model for our predictor. The output of the predictor using the decision tree model is shown in Fig. 5(b). Its accuracy is 72%. The R-square of the predictor is 0.2747, while the RMSE is 0.1818. We can see that the predictor using a decision tree model has a worse result than the predictor using a linear regression model does.

Random Forest. We now explore a predictor using the random forest model. The maximum depth of the random tree regressor is set to eight, and the amount of estimators is set to 100. The output of the predictor using the random forest is shown in Fig. 5(c). Its accuracy is 89%. The R-square of the predictor is 0.3941, while the RMSE is 0.1662. It can be seen that the predictor using a random forest model is more suitable for the parking behavior prediction.

We then show the ranking of features using random forest model in Fig. 5(d). We can see that `season` is the most important features, and `weathersit`, which relates to precipitation, is also important to the model. Among the numerical features, the feature `temperature` has the most significant impact on the target variable `cnt`.

6.3 Results

As presented in Sect. 6.2, we have implemented five regression models (three linear regression models, a decision tree model, and a random forest model) for park behavior prediction. The decision tree model gives the worst result; its accuracy is only 72%. The linear regression model achieves an accuracy of 78%. The random forest model presents the best result; its accuracy is 89%.

7 Related Work

Urban Freight Parking Demand Prediction. Alho and others [3] proposed a prediction method for urban freight parking demand using ordinary least squares (OLS) linear regression and generalized linear models (GZLMs). This work helps parking lot managers to prediction the demand for parking space for freight cars.

Prediction of Parking Space Availability. Parking space availability prediction [18,21,30] is an indispensable part for intelligent parking system. Caicedo and others [5] proposed a method for predicting space availability in an IPR architecture for parking facility information systems.

Prediction of Parking Space Occupancy. Pierce and others [20] proposed a framework, SFpark, aiming to solve the problems created by charging too much or too little for curb parking. Simhon and others [29] extended SFpark with a machine learning approach for better prediction. Chen [6] studied parking occupancy prediction and pattern analysis. Hossinger and others [11] developed a real-time occupancy model of short-term parking zones. Florian and others [8] presented a model for predicting parking occupation.

Influencing factors of Parking Space Usage. There are many works about influencing factors of parking space usage, including pricing strategy, traffic condition, and parking lot locations. Pierce and others [19] provided an evaluation of pricing parking by demand. Ottosson and others [17] studied the sensitivity of on-street parking demand in response to price changes. Lam and others [13] proposed a bilevel programming model to determine the minimum supply of parking spaces.

8 Conclusion

In this paper, we have analyzed parking behavior with weather conditions considered. We exhibit our method about how to perform preprocessing and feature selection from data, and also explore different regression models for parking behavior prediction. Experiments show that the random forest model has the best results, which achieves 89% accuracy.

Acknowledgments. This work is partially supported by the National Key R&D Program of China (Grant No. 2017YFB1003103), National Natural Science Foundation of China (Grant No. 61722208, 61732014, 61802412). Feng Zhang is the corresponding author (fengzhang@ruc.edu.cn).

References

1. Abdullatif, A., Masulli, F., Rovetta, S.: Tracking time evolving data streams for short-term traffic forecasting. Data Sci. Eng. **2**(3), 210–223 (2017)
2. Aggarwal, C.C.: Outlier Analysis. Data Mining, pp. 237–263. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-14142-8_8
3. Alho, A.R., Silva, J.D.A.E.: Freight-trip generation model: predicting urban freight weekly parking demand from retail establishment characteristics. Transp. Res. Rec. **2411**(1), 45–54 (2014)
4. Banti, K., Louta, M., Karetsos, G.: ParkCar: a smart roadside parking application exploiting the mobile crowdsensing paradigm. In: 2017 8th International Conference on Information, Intelligence, Systems & Applications (IISA), pp. 1–6. IEEE (2017)
5. Caicedo, F., Blazquez, C., Miranda, P.: Prediction of parking space availability in real time. Expert Syst. Appl. **39**(8), 7281–7290 (2012)
6. Chen, X.: Parking occupancy prediction and pattern analysis. Dept. Comput. Sci., Stanford Univ., Stanford, CA, USA, Technical Report CS229-2014 (2014)
7. Fang, J., Ma, A., Fan, H., Cai, M., Song, S.: Research on smart parking guidance and parking recommendation algorithm. In: 2017 8th IEEE International Conference on Software Engineering and Service Science (ICSESS), pp. 209–212. IEEE (2017)
8. Florian, M., Los, M.: Impact of the supply of parking spaces on parking lot choice. Transp. Res. Part B: Methodol. **14**(1–2), 155–163 (1980)
9. Girden, E.R.: ANOVA: Repeated Measures, No. 84. Sage, Thousand Oaks (1992)
10. Hans, C.: Bayesian lasso regression. Biometrika **96**(4), 835–845 (2009)
11. Hössinger, R., Widhalm, P., Ulm, M., Heimbuchner, K., Wolf, E., Apel, R., Uhlmann, T.: Development of a real-time model of the occupancy of short-term parking zones. Int. J. Intell. Transp. Syst. Res. **12**(2), 37–47 (2014)
12. Kong, D., Li, F., Zhang, B.: Design and implementation of intelligent management system for urban road parking. In: Journal of Physics: Conference Series, vol. 1087, p. 062061. IOP Publishing (2018)
13. Lam, W.H., Tam, M., Yang, H., Wong, S.: Balance of demand and supply of parking spaces. In: 14th International Symposium on Transportation and Traffic Theory Transportation Research Institute (1999)
14. Le Cessie, S., Van Houwelingen, J.C.: Ridge estimators in logistic regression. J. Roy. Stat. Soc.: Ser. C (Appl. Stat.) **41**(1), 191–201 (1992)

15. Liaw, A., Wiener, M., et al.: Classification and regression by randomForest. R. News **2**(3), 18–22 (2002)
16. Liu, L., Yang, S., Peng, L., Li, X.: Hierarchical hybrid memory management in OS for tiered memory systems. IEEE Trans. Parallel Distrib. Syst. (2019)
17. Ottosson, D.B., Chen, C., Wang, T., Lin, H.: The sensitivity of on-street parking demand in response to price changes: a case study in Seattle, WA. Transp. Policy **25**, 222–232 (2013)
18. Pflügler, C., Köhn, T., Schreieck, M., Wiesche, M., Krcmar, H.: Predicting the availability of parking spaces with publicly available data. Informatik **2016** (2016)
19. Pierce, G., Shoup, D.: Getting the prices right: an evaluation of pricing parking by demand in San Francisco. J. Am. Plann. Assoc. **79**(1), 67–81 (2013)
20. Pierce, G., Shoup, D.: SFpark: pricing parking by demand (2013)
21. Quinn, J.: System and method for predicting parking spot availability, February 28 2008. US Patent App. 11/849,493
22. Rasmussen, C.E.: Gaussian processes in machine learning. In: Bousquet, O., von Luxburg, U., Rätsch, G. (eds.) ML-2003. LNCS (LNAI), vol. 3176, pp. 63–71. Springer, Heidelberg (2004). https://doi.org/10.1007/978-3-540-28650-9_4
23. Roman, C., Liao, R., Ball, P., Ou, S., de Heaver, M.: Detecting on-street parking spaces in smart cities: performance evaluation of fixed and mobile sensing systems. IEEE Trans. Intell. Transp. Syst. **19**(7), 2234–2245 (2018)
24. Safavian, S.R., Landgrebe, D.: A survey of decision tree classifier methodology. IEEE Trans. Syst. Man Cybern. **21**(3), 660–674 (1991)
25. Seabold, S., Perktold, J.: Statsmodels: econometric and statistical modeling with python. In: 9th Python in Science Conference (2010)
26. Seber, G.A., Lee, A.J.: Linear Regression Analysis, vol. 329. Wiley, Hoboken (2012)
27. Shahzad, A., Choi, J.Y., Xiong, N., Kim, Y.G., Lee, M.: Centralized connectivity for multiwireless edge computing and cellular platform: a smart vehicle parking system. Wirel. Commun. Mob. Comput. **2018**, 1–23 (2018)
28. Shin, J.H., Kim, N., Jun, H.b., Kim, D.Y.: A dynamic information-based parking guidance for megacities considering both public and private parking. J. Adv. Transp. **2017**, 1–19 (2017)
29. Simhon, E., Liao, C., Starobinski, D.: Smart parking pricing: A machine learning approach. In: 2017 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), pp. 641–646. IEEE (2017)
30. Tayade, Y., Patil, M.: Advance prediction of parking space availability and other facilities for car parks in smart cities. Int. Res. J. Eng. Technol. **3**(5), 2225–2228 (2016)
31. Tilahun, S.L., Di Marzo Serugendo, G.: Cooperative multiagent system for parking availability prediction based on time varying dynamic markov chains. J. Adv. Transp. **2017**, 1–14 (2017)
32. Zhang, F., et al.: An adaptive breadth-first search algorithm on integrated architectures. J. Supercomput. **74**(11), 6135–6155 (2018)