Predicting Car Park Occupancy Rates in Smart Cities

Daniel H. Stolfi[1], Enrique Alba[1], and Xin Yao[2]

[1] Departamento de Lenguajes y Ciencias de la Computación,
University of Malaga, Malaga, Spain {dhstolfi, eat}@lcc.uma.es
[2] CERCIA, School of Computer Science, University of Birmingham, Birmingham,
U.K. x.yao@cs.bham.ac.uk

Abstract. In this article we address the study of parking occupancy data published by the Birmingham city council with the aim of testing several prediction strategies (polynomial fitting, Fourier series, k-means clustering, and time series) and analyzing their results. We have used cross validation to train the predictors and then tested them on unseen occupancy data. Additionally, we present a web page prototype to visualize the current and historical parking data on a map, allowing users to consult the occupancy rate forecast to satisfy their parking needs up to one day in advance. We think that the combination of accurate intelligent techniques plus final user services for citizens is the direction to follow for knowledge-based real smart cities.

Keywords: Smart city, smart mobility, parking, K-means, time series, machine learning

1 Introduction and Related Works

Finding an available parking space is hard in most big cities, especially in the city center. Off-street car parks are a viable alternative, especially when the number of inhabitants in urban areas is increasing and expected to rise to 75% of the world's population by 2050 [1]. On-street parking spaces are quite limited and usually it is cheaper to find an off-street car park or pay and display bays rather than wasting time (and fuel) in finding a free space. Not to mention the health consequences [2] provoked by an increase of not only air pollution but also drivers' stress. However, even paid spaces are scarce nowadays as city infrastructures have not grown in line with population growth.

Fortunately, smart city initiatives are changing this [3]. One of the main aspects of a smart city is the so-called Internet of Things (IoT). The main idea is to know the state of a city by using sensors to monitor such data as the road traffic state, temperatures, pollution levels, and car parks' occupancy rates. Although monitoring single parking spaces is not economically viable, it is possible to count the number of vehicles entering and leaving an off-street car park and make these data publicly available to help make decisions (and predictions) based on them.

The prediction of car park availability is the subject that has been studied in a context of smart cities, especially now when most parking facilities have installed sensors as part of their infrastructure.

In [4], the authors fit a continuous-time Markov model to predict future occupancies in several parking locations to propose different alternatives to drivers. They consider not only the car park occupancy rate but also the estimated time of arrival obtained from the vehicle's navigation system in which the calculations are done. They provide two *ad hoc* examples to test their proposal, showing promising results. In this article we take a different approach where, instead of using a navigator, any Internet capable device, such as a mobile phone, can be used to check the current/future state of the desired car park.

Two smart car park scenarios based on real-time information are presented in [5]. The authors use historical data made available by the authorities of the cities of San Francisco, USA and Melbourne, Australia. They employ Regression Tree, Neural Networks and Support Vector Regression as prediction mechanisms for the parking occupancy rate. Their experiments reveal that the regression tree using the historical data in combination with times and weekdays, performs best for predicting parking availability on both data sets. We have analyzed different predictors in our analysis, however, it would be interesting to compare our results to those produced by these alternative predictors in the future.

In [6] a methodology for predicting parking space availability in Intelligent Parking Reservation architectures is proposed. It consists of a real-time availability forecast algorithm which evaluates each parking request and uses an aggregated approach to iteratively allocate parking requests according to drivers' preferences, and parking availability. They employ historical information of entering and leaving to update and predict the availability for each parking alternative. The results provided, obtained from contrasting predictions with real data, show that the forecast is adequate for potential distribution in real-time. Our approach differs from this proposal in that we study different predictors and do not interact with the current demand, relying just on the historical data.

In short, our proposal consists in studying the different prediction strategies to analyze the historical occupancy rates of car parks and forecast the future availability, presenting this information to the users in a web page.

The rest of this paper is organized as follows: Section 2 describes the system architecture, including the web page and the predictors. In Section 3 we discuss the predictor techniques, the training and testing stages, and our results. Finally, in Section 4, conclusions and future work are presented.

2 System Architecture

The architecture of our system (Fig. 1) comprises Downloaders, the Data Parser, the Database, the Predictor and the Web Page in which both, the current state of the car parks and the predictions made, are presented to the public.

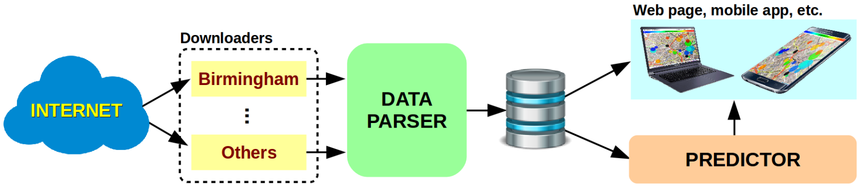


Fig. 1. Schema of the data parser.

2.1 Downloaders

These modules access different data sources available from the Internet to obtain the occupancy data of the car parks defined in the system. They are set up with the source URL, the frequency of readings, as this has to be adapted according to each data source, and the possible data transformations (CSV, XML, etc.) to be completed before feeding the Data Parser. Note that in this study we are working with just one data source.

2.2 Data Parser

The Data Parser processes the data provided by the Downloaders and stores them in the Database. It also checks the validity of the car parks, creating new ones if necessary, whilst avoiding data redundancies.

2.3 Database

The Database stores the data collected from each car park so that it can be shown on the web page. We store the code, description, capacity, latitude and longitude of each car park, as well as the city to which it belongs. Periodically, we also store occupancy data for car parks consisting of spaces used, state, and last updated time.

Additionally, the historical data of the car parks is obtained by the Predictor from the database to be used for forecasting their future occupancy.

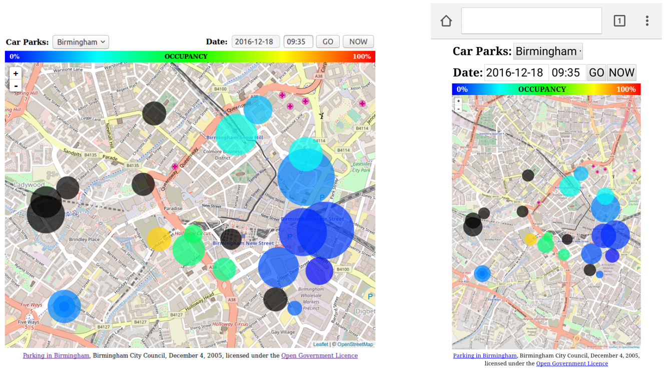
2.4 Web Prototype

Data stored in the database can be shown at any time in our web page and mobile app. We present each car park geolocated in its real geographical position in the map by using the library Leaflet[3] and the tiles from OpenStreetMap[4].

Figure 2 shows snapshots of the web page as visualized on a computer desktop and on a mobile phone. We can see that each car park is shown as a circle whose size is proportional to the number of parking spaces and its color represents the occupancy rate as shown in the upper scale, i.e. blue for totally free and red for full. Car parks whose data is out of date are shown in black.

[3] http://leafletjs.com/
[4] http://www.openstreetmap.org/

Additionally, the user can choose different dates and hours to see the historical data and it is possible to select future dates to obtain an occupancy prediction, as well.


Car Parks: Birmingham

Date: 2016-12-18 09:35 GO NOW

Parking in Birmingham, Birmingham City Council, December 4, 2005, licensed under the Open Government Licence

(a) Desktop web page.

(b) Mobile version.

Fig. 2. Web page and mobile prototypes presenting the geolocation, state, capacity, and occupancy of each car park.

2.5 Predictor

Data stored in the database is also used to predict future occupancy of the car parks. We have experimented with six different predictors (Fig. 3): Polynomial Fitting [7], Fourier Series [8], K-Means [9], KM-Polynomials, Shift & Phase, and Time Series [10] which are all described in the next section. We have selected them for this initial study because they are simple, easy to implement, and they allow us to represent each car park with just a few parameters. Furthermore, they are present in the open data provided by cities nowadays.

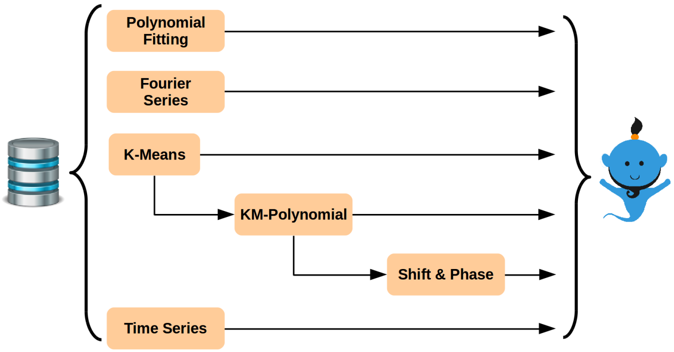


Fig. 3. Different predictors tested and the existing relationships between some of them.

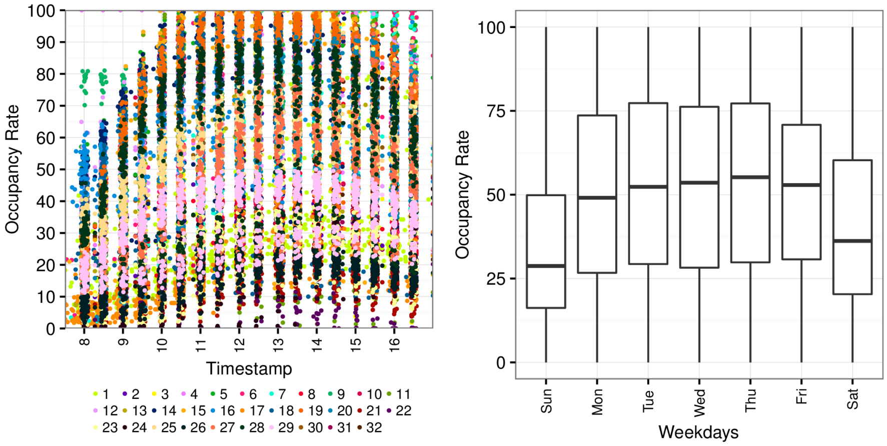
3 Prediction Techniques

We wish to address the prediction of the future occupancy rates of car parks in a city. For our prototype we have chosen the data set "Parking in Birmingham" published by Birmingham City Council[5] in the United Kingdom, licensed under the Open Government License v3.0. It includes the car parks operated by NCP (National Car Parks) in that city, and is updated every 30 minutes from 8 AM to 5 PM. In our study, we worked with data from Oct 4th 2016 to Dec 19th 2016 (11 weeks).

The data provided is not very accurate as sometimes the sensors are faulty or even, the whole data set may not be updated for a whole day. To address these situations we implemented a filtering stage before the data processing as follows:

1. The occupancy rate is calculated based on each car park's capacity.
2. The percentage values beyond the range (0-100%) are adjusted to these limits.
3. Out of date data is discarded.
4. If the variability of a car park's occupancy for an entire day is below 5% it is assumed that it is due to faulty sensors and that day is discarded.
5. Data on a car park that is below 5% for an entire day is also discarded.
6. Car parks without data are excluded from the study.

Figure 4(a) shows the occupancy data available for all the car parks and dates after filtering the initial data set and in Fig. 4(b) the data distribution of the car park occupancy on weekdays is depicted as a boxplot.



(a) Occupancy data after filtering.

(b) Distribution along weekdays.

Fig. 4. Occupancy data from the 32 car parks and their distribution on weekdays.

[5] https://data.birmingham.gov.uk/dataset/birmingham-parking

We can see that occupancy rates decreases on Saturdays and Sundays as expected, while they are quite similar throughout the rest of the week. All in all, we finally had a working data set consisting of 32 car parks and 36,285 occupancy measures.

Throughout our study we used the Mean Squared Error (MSE) to test the quality of the predictions made, not only in the training stage but also in the testing stage. Equation 1 presents the MSE formula where y_i are the measured real values, f_i are the fitted ones, and n is the number of observations.

$$MSE = \frac{1}{n} \sum_i (y_i - f_i)^2 \tag{1}$$

3.1 Training

Prior to training, data ought to be processed so as to guarantee a fair comparison between the different predictors used. We wished to predict the occupancy rate of each car park over an entire week, consequently, we decided to use a different predictor for each car park and weekday to conduct this first initial study. We selected the first ten weeks of data (Oct 4th to Dec 12th) for training and left the eleventh week for the testing stage to simulate the real use of the web page by a user.

As we have pointed out sometimes sensors fail. To address this, first, we discarded data from a car park for an entire day when more than 25% of the measures were missing. Second, if a car park did not have at least one weekday of training data, it was also excluded.

After applying this second filter to the training data set, we finished with 29 car parks which presented reliable occupancy data. However, not all of them had the same number of occupancy measures, as our filter is not very restrictive.

In order to achieve a fair comparison, especially for the Time Series predictor, we completed the training data by i) adding non-existent measures by repeating the previous value, i.e. if there was no data at 11:30 AM we therefore created a measure with the same value as 11:00 AM (Birmingham car parks update every half hour); and ii) duplicating the previous weekday if an entire day was missing, i.e. if data from Tuesday 8th was missing we created the occupancy data by copying the values from Tuesday 1st. Note that this was checked for each car park as we completed each one, individually. After completing the data, our training data set involved occupancy values for 29 car parks over ten days from 8:00 AM to 4:30 PM.

Additionally, as each predictor has its own trade-off between accuracy and number of parameters, we performed a parameterization and selected the ones that best suited to our study by using the *elbow* method [11]. This is a visual method to obtain the most promising value from a line chart where a change in the slope looks like the elbow of an arm.

To improve the training process we decided to use K-Folds cross validation. We used 10 sets (we have 10 weeks of data for training) where each training set consisted of 32,886 occupancy data values (29 car parks, 9 weeks, 7 days, 18 values per day). To obtain the average MSE values we tested each predictor on the remaining week (3,654 values). By selecting a different test week we obtained ten different training and testing sets to train our predictors as discussed below.

Polynomial Fitting (P) This predictor consists in a polynomial fitted to each car park and weekday. We studied different polynomial degrees to find which value presented a low average MSE. We also wished to keep a reduced number of parameters to represent each car park and weekday. Figure 5(a) presents the MSE values obtained after using cross-validation for all ten training processes for Birmingham. We can see that according to the *elbow* method, polynomials of second degree are the best choice to be used in this predictor for the ten cases because they have only a few parameters and good precision.

Fourier Series This predictor consists in fitting a Fourier series to each car park and weekday. In this case we considered different numbers of components. Since they are composed of pairs of sines and cosines, the different alternatives tested are always odd numbers (a constant component is included, as well). In Fig. 5(b) the MSE values obtained after using cross-validation to train the Fourier predictor are depicted. The *elbow* method clearly states that using just three components (a constant, a sine, and a cosine) is the best choice in all cases, not only because of the change in the slope, but also because it leads to low MSE values without increasing the number of components.

***K*-Means** Clustering by using K-Means is a method that allows grouping pairs of car parks and weekdays in different clusters whose centroid represents the whole set of occupancy measures in the group. It is an interesting way of describing a set of car parks which behave similarly. We tested up to ten clusters to decide which option was better according to the MSE values and the *elbow* method. Figure 5(c) presents the MSE values obtained for each fold and number of clusters. It can be seen that three clusters is a good choice for this predictor for all the training folds.

KM-Polynomials This predictor fits a polynomial to the existing centroid points of each cluster calculated by K-Means. This step is necessary to improve the accuracy of the predictions by interpolating a polynomial to the points in each centroid as they are spaced according to the frequency of the measures, i.e. 30 minutes. In Fig. 5(d) the MSE values obtained for different polynomial degrees are depicted. We can see that, using the *elbow* method, the best degree of the polynomials matches the one selected for the Polynomial Fitting predictor. This was somewhat expected, as the centroids ought to represent a set of measures which are the same as the ones used to obtain the aforementioned polynomials.

Shift & Phase (SP) To improve the accuracy of the prediction even further, we defined a new predictor which uses the KM-Polynomials calculated in the previous section and adds two new parameters (δ and ϕ) in order to modify the shift (y axis) and the phase (x axis) of the original polynomial as shown in Equation 2.

$$F(x) = (a_0 + \delta) + a_1 \times (x + \phi) + a_2 \times (x + \phi)^2 + \ldots + a_n \times (x + \phi)^n \qquad (2)$$

Then, by using a weighted nonlinear least-squares estimation [12], the values for δ and ϕ for each pair of car park and weekday were obtained, so that a car park's occupancy rate could be predicted by following the process shown in Algorithm 1.

Algorithm 1 Occupancy Prediction.

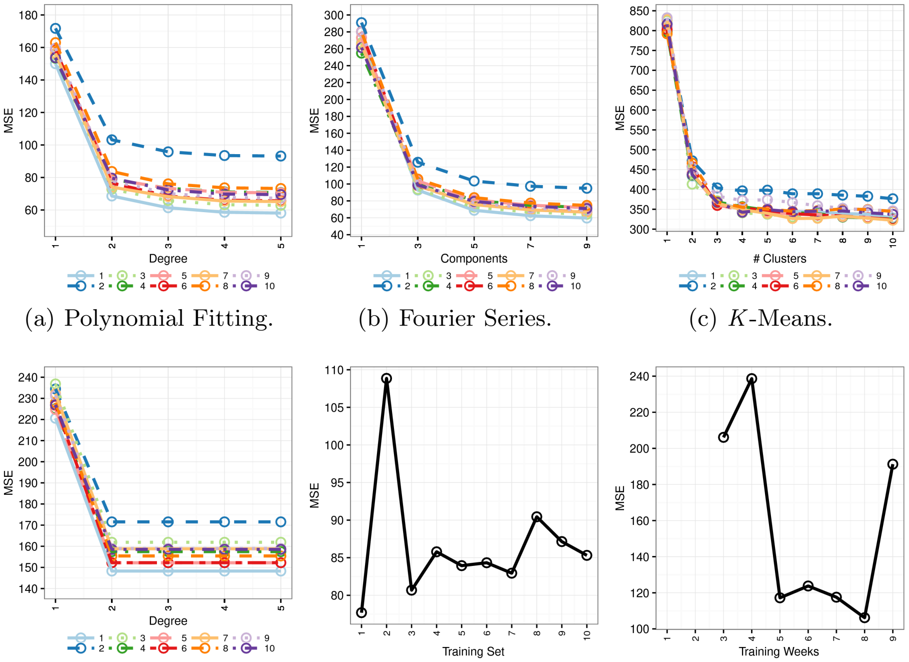
```
function OccupancyPrediction($car\_park, wd, time$)
    $cluster\_id \leftarrow getClusterId(car\_park, wd)$
    $coefs \leftarrow getPolynomialFitting(cluster\_id)$
    $(\delta, \phi) \leftarrow getShiftPhase(car\_park, wd)$
    $occupancy \leftarrow getOccupancy(time, coefs, \delta, \phi)$
    return $occupancy$
end function
```

The main function receives as parameters the *car_park* identity, the weekday (*wd*), and the *time* at which we want to know the occupancy. Inside the function, the corresponding *cluster_id* is obtained based on the *car_park* and the weekday *wd*, as the first step. Second, the coefficients *coefs* of the polynomial fitted to the cluster's centroid are obtained. Third, δ and ϕ for the car park and weekday are also obtained. Finally, the *occupancy* value is calculated by using the formula in Equation 2 whereas x is the *time* parameter.

Figure 5(e) shows the best data set according to the MSE values obtained when training SP. After all these experiments, it is clear that data set 1 (when we train with 2 to 9 and test on 1) presented the best results, i.e. the lower average MSE values for all the predictors trained by using cross validation.

Time Series (TS) To train the Time Series (TS) predictor a different approach was followed, as a consecutive, ordered number of time periods (weekdays) are needed, which makes it impossible to use k-folds. We therefore trained a different time series for each car park and weekday to be consistent with the other predictors analyzed here.

Figure 5(f) shows the MSE values obtained when training the TS predictor with different numbers of weeks. It is worth noting that having more data did not imply computing the best prediction according to our experiments. Nevertheless, it was something to be analyzed at a later date, as there were not enough data to make a solid conclusion at that point. Furthermore, a variation in the test week such as a bank holiday may also increase the MSE value.



(a) Polynomial Fitting. (b) Fourier Series. (c) *K*-Means.

(d) KM-Polynomials. (e) Shift & Phase. (f) Time Series.

Fig. 5. Average MSE values obtained when training the different predictors by using k-fold cross validation (10 folds). Figures 5(a) to 5(d) also shows the parameterization performed.

3.2 Prediction

In this next step we compare the predictions made by the predictors trained in the previous sections. To do so we predicted the occupancy rates for seven unseen days (from Dec 13th to Dec 19th) and compared the MSE values obtained for each region. We did not complete this data set as we did in the training stage as isolated values are also useful to test our predictions providing they are produced by reliable sensors. All in all, we have tested our predictor on 3,425 occupancy values: 480 for Sunday, 468 for Monday, 493 for Tuesday, 493 for Wednesday, 501 for Thursday, 522 for Friday, and 468 for Saturday.

Figure 6 shows the boxplots of the distribution of the MSE values for each predictor. We can see that TS performed best for weekdays followed by P and SP which show very good results for the whole week except Mondays. KM and KP are the worst predictors as they are based on just the three clusters defined in training. Fourier improved upon KM and KP on all weekdays except on Mondays which was clearly the hardest day for the predictors, followed by the weekend. Saturdays and Sundays are the days when people do not follow a clear pattern of behavior as they do on working days, which, in part, explains the larger MSE values observed. On the other hand, the occurrence of large MSE values observed for Monday has to be further investigated as they clearly differ from other days.

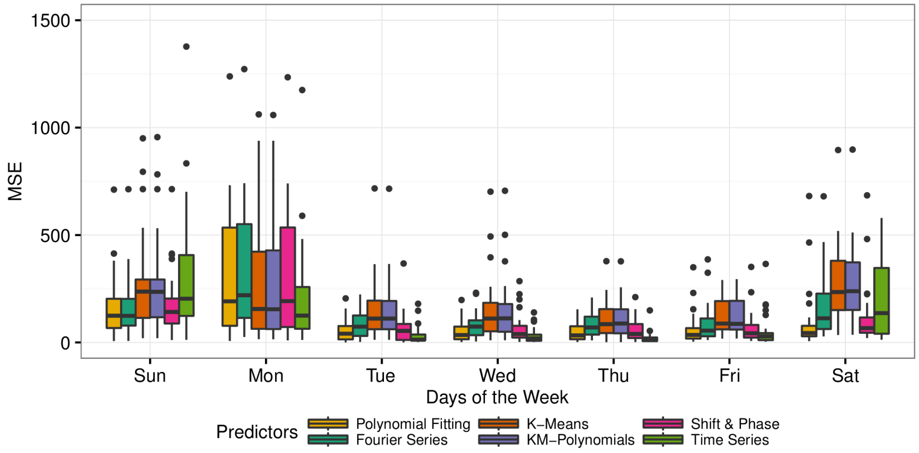


Fig. 6. Average MSE of each predictor by weekdays.

4 Conclusions and Future Work

In this article we have presented six predictors for forecasting car park occupancy rates in the city of Birmingham. We have trained them by using real data published by the local council and presented the results obtained after testing them with one week of unseen parking data.

Although there is no clear winner, the time series predictor seems to show the best results. Shift & Phase also has good results, especially if we take into account that it is simpler and requires just five parameters to predict a car park's occupancy rate instead of a series of values.

Our proposal is a novel service as although there are web pages offering information on car park's occupancy rates, they rarely make predictions of the next day's state.

As a matter for future work we wish to develop a mobile app, repeat this study using a larger training data set, and include new predictors in the comparison, e.g. a multivariate predictor. Additionally, we wish to address another method for missing values such as average, previous days.

5 Acknowledgements

This research is partially funded by the Spanish MINECO project TIN2014-57341-R (http://moveon.lcc.uma.es). Daniel H. Stolfi is supported by a FPU grant (FPU13/00954) from the Spanish Ministry of Education, Culture and Sports. University of Malaga. International Campus of Excellence Andalucia TECH.

References

1. Bakici, T., Almirall, E., Wareham, J.: A Smart City Initiative: the Case of Barcelona. Journal of the Knowledge Economy **4**(2) (2013) 135–148
2. Hertel, O., Jensen, S.S., Hvidberg, M., Ketzel, M., Berkowicz, R., Palmgren, F., Wåhlin, P., Glasius, M., Loft, S., Vinzents, P., Raaschou-Nielsen, O., Sørensen, M., Bak, H.: Assessing the Impacts of Traffic Air Pollution on Human Exposure and Health. In: Road Pricing, the Economy and the Environment. Advances in Spatial Science. Springer Berlin Heidelberg (2008) 277–299
3. Neirotti, P., De Marco, A., Cagliano, A.C., Mangano, G., Scorrano, F.: Current trends in Smart City initiatives: Some stylised facts. Cities **38** (jun 2014) 25–36
4. Klappenecker, A., Lee, H., Welch, J.L.: Finding available parking spaces made easy. Ad Hoc Networks **12**(1) (2014) 243–249
5. Zheng, Y., Rajasegarar, S., Leckie, C.: Parking availability prediction for sensor-enabled car parks in smart cities. In: 2015 IEEE Tenth International Conference on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP). (2015) 1–6
6. Caicedo, F., Blazquez, C., Miranda, P.: Prediction of parking space availability in real time. Expert Systems with Applications **39**(8) (2012) 7281–7290
7. Fan, J., Gijbels, I.: Local polynomial modelling and its applications: monographs on statistics and applied probability 66. Volume 66. CRC Press (1996)
8. Butzer, P.L., Nessel, R.J.: Fourier analysis and approximation. Volume 40. Academic Press (2011)
9. Hartigan, J.A., Hartigan, J.A.: Clustering algorithms. Volume 209. Wiley New York (1975)
10. Fuller, W.A.: Introduction to statistical time series. Volume 428. John Wiley & Sons (2009)
11. Sugar, C.A.: Techniques for clustering and classification with applications to medical problems. PhD thesis, Stanford University (1998)
12. Draper, N.R., Smith, H., Pownell, E.: Applied regression analysis. Volume 3. Wiley New York (1966)