A mathematical model for the occupation rate in a neighborhood

Clarisha Nijman

February 5th, 2019

Preface

The growth in cars in the Netherlands have an impact on the time spent to find a parking space and on the quality of the traffic. ARS Traffic & Transport Technology (ARS T & TT) is a company in the Netherlands that is interested in solving such problems. This company focuses on Traffic planning systems and monitoring and operation of Intelligent Transportation System (ITS) solutions by developing software systems on both national and international scale, to make mobility smarter, faster, safer and more convenient.

As an OR student at the University of Twente, the main goal for this project was to find a model to predict the occupation rate of the parking place in a neighborhood. Such a model will help design software to inform drivers of the free parking spaces in a neighborhood at a point in the future. Instead of continuing to cruise for parking, a driver can then opt to look for a parking space in a neighborhood, a parking garage or a Park and Ride area.

The report resulting from this research is entitled: "A mathematical model to predict the occupancy rate of the parking place in a residential area". Doing this assignment has given me much more insight into the use of mathematical models, and the application of the mathematical concepts such as the binomial distribution, the convolution, the convex function and Markov chains. Furthermore, my knowledge about the R and the MATLAB software has increased.

Without the help of primarily the academic mentor, the business mentor, school mates and many others, this project could not be carried out properly. I would therefore like to thank the academic mentor, dr. J.C.W.van Ommeren, for being patience and calm. I really appreciate his watchful eyes and especially his valuable feedback and the space he offered me to be myself although within a "limited area". Furthermore, my thanks also go to Okialmasamalia, MSc. Jaap Slotenbeek and the online MATLAB crew, who helped me to get along with the MATLAB code.

For successfully finishing this master course, I owe also many thanks to: the study advisor Ms. L. Spijkers and the guardian of foreign students drs J. Schut, my mentor prof.dr. R.J. Boucherie, the director of the program dr. J.W. Polderman and many other teachers and staff members. For the external support I want to thank my mother and her husband for the good care here in the Netherlands and the brothers and sister of congregation Mekkeltholt of Jehovah's Witnesses for the many encouraging words.

My thanks also go to AdeKUS who made the training possible, in particular dr. S. Venetian, drs. H. Antonius and drs. C. Gorison together with J. Simons-Turney, W. Valies and drs. R. Peneux who arranged that I could get permission from the ministry to do this study outside Suriname. Furthermore, my thanks go to my sister and my father for taking care of business affairs in Suriname.

Contents

Abstract

Keywords: *ARIMA models, data analysis with Markov chains, high order Markov chain models, occupancy rate, parking analysis, parking conventions, parking demand, parking modeling, parking policy, parking problems in the Netherlands, parking research, prediction models for Markov chains.*

Now a days both the population and the number of cars in the Netherlands is growing fast such that finding an empty parking space is hard. Lack of enough parking spaces leads to cruising, a time-consuming phenomenon that is bad for the environment and also for the health of people. Digital information about the number of empty parking space close by would be helpful for drivers especially during rush hour.

Therefore ARS TT&T wants a model to predict the occupancy rate in a neighborhood such that information could be given to drivers who are looking for a single parking space. The main question of this research is about: *To what extend can a Markov chain prediction model be used to predict the distribution of the occupancy rate of a parking lot in a neighborhood based on the ARS data files?* This question was explored based on the following sub questions: How important is knowledge about the distribution of parking times for visitors and for permit holders? What is the optimum fraction of parking spaces that should be equipped with a sensor? What is the sensitivity of the fraction of with a sensor-equipped parking space? What is the sensitivity of the number of scans per day and the distribution of the scans over the day? Are there other data sources that can provide extra information?

The number of cars for every minute between 9.00am and 21.00pm for 500 days on PARK200 is deduced from the data. Each minute a single parking space can be either empty or not. As it is not clear what happens with parkers at the last minute of the day it is assumed that these cars stay overnight such that the parking time of these cars is at least 720 minutes. The short- and long-term parkers are found with the distribution of the parking time.

The parking process can be described as a two-dimensional Markov process with Poisson arrivals, general service or parking time, c servers or parking spaces and maximum c cars in the system. An important assumption in this process is that parkers do decide independent from each other how long they will stay at the parking place. This idea suggests that the short-term parkers in the system only influences the maximum number of long-term parkers that can enter the system at time t. The actual number of cars that enters the system depends on the parking demand and the available parking space.

The situation at the parking place can be modeled as a non-homogeneous two-dimensional Markov chain. Predictions were done for each dimension separately with the first and higher order Markov chain prediction model. The transition probabilities were determined with the arrival-departure behavior and with the fit distribution of the transitions. The non-homogeneity of the chain was tackled by estimating the transition probabilities with data coming from a time interval containing time t. In this time interval it is assumed that the Markov chain is homogeneous.

The research reveals that the higher order models as proposed by Chin was the best mathematical model in combination with some mathematical techniques. These techniques do take care of the two-dimensionality of the process and the non-homogeneity of the chain. There were also mathematical techniques used to correct for prediction flaws.

This report starts with a section that describes the magnitude of the parking problem, followed by the problem description and a discussion of the research variables. Section two zooms in on the data sets. The next section addresses the assumptions and restrictions needed to make this study operational, followed by a mathematical problem description. Section 4 contains the mathematical concepts used in this report and section 5 a discussion of the way the model will be applied together with techniques. The next section in this report highlights some interesting results. The last section in this report regards conclusions and recommendations.

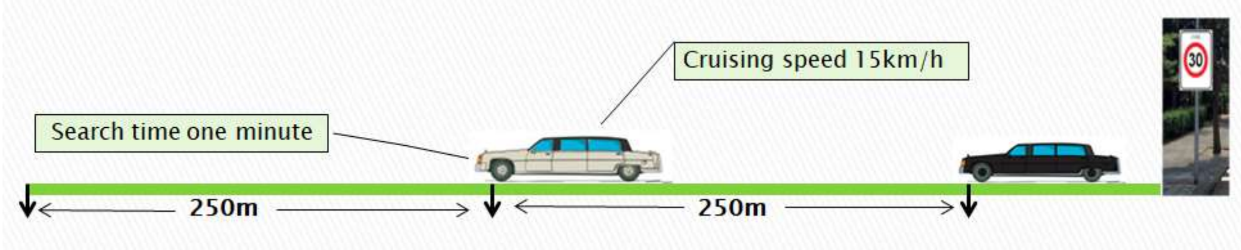
Introduction to the Parking Problems

The year 1958 is characterized as the beginning of the mass motorization in the Netherlands, or the starting period of spectacular growth of the number of cars in this country. From then on, municipalities have also implemented a parking policy and in the 1970's municipalities even were "obliged" to write a parking plan (*Stienstra, 2011, p7*). Now decades later, the increase in cars is still noticeable in the Netherlands. At the start of 2016, the Netherlands had almost 7.2 million private passenger cars, almost 900 thousand more than ten years earlier. This growth is 1.125 time the population of 18 years and older, which grew by more than 800 thousand people in that same period. Car ownership also increased from 494 cars per thousand inhabitants in early 2006 to 530 in early 2016 (*CBS, 2017, p 7*).

With this increase in cars, the need arises to place or park cars somewhere, whether people take their car to relocate or not (*CBS, 2017, page 7*). This growth therefore has far-reaching consequences for the organization of the country. Each one of the millions of cars in the Netherlands is parked somewhere on average 23 hours a day. Cars are used to travel between home, the office, the shopping center, the sports field or many other locations. In fact, compared with the number of cars, twice as much parking spaces are needed to meet this parking demand (*CROWS Ede, 2014*). Meijer *(2018)* stated that cars are parked on average 95% (22.8 hours) of the day, and in case a person possesses a second car, that percentage is 99% (23.8 hours).

If there is no proper response to the demand of parking spaces, there might be an increase in cruising in order to park. In large cities, the effect of cruising is particularly noticeable during rush hours. Studies have shown that 8 percent to 74 percent of the traffic flow is cruising for parking (*Shoup, 2006*). Using data generated by Dutch National Travel Survey (MON) for the years 2005–2007 it was proven that 30% of the car drivers cruise before finding a parking spot, and most of this group cruised for one minute (*Van Ommeren et, 2012*). According to Gantelet (2006) the average car parking search time in three French cities (Grenoble, Lyon, Paris) is around 8.4 minutes. Another observation is the high variability of the search time for one occupancy ratio value, especially when the latter is higher than 85% (*Belloche, 2015, p 6, 313-324*). This implies larger search times when the demand for parking is high.

Take for example a realistic scenario in Amsterdam to illustrate the congestion this could create for the traffic. Suppose that a car starts cruising at a road where the allowed speed of traffic equals 30km per hour. With a cruising speed of 15km per hour and a search time of one minute, it is expected to find a parking spot after 250 meters. This car will not hinder a next car behind him at a minimum distance of 250m when starting the search. But how realistic is it that the distance between two cars driving on a road in Amsterdam equals 250 meters? According to the yearbook 2017, this city counts 231,183 cars and a total road length of 1710 km under the management of the municipality (*OIS, 2017b, p112, 114*). That implies a ratio of 135 cars per km, and even if 90% of the cars are parked somewhere it means 3.4 cars per 250m road length. This scenario pictures how easily a driver that starts to cruise might affect at least 2 cars driving after him with a speed of 30 km per hour.



Add to this the effect of the 15.7 million visitors of Amsterdam in 2016 (*CBS, 2018*). More than half of these visitors, 51%, used a car to go from one place to another. Amsterdam's tourists also relocate 6.5 times a day on average (*OIS, 2017*). Cruising can contribute to congestion especially during peak hours. According to the INRIX (2018) the average time spent in peak congestion is 5.5 minutes for cities in the Netherlands (*INRIX, 2018 p13)*.

Cruising for parking is time consuming but costs also money and deteriorates our environment. Shoup (2005) conducted a 'cruising for parking' study in the Westwood village, a commercial district bordered by the UCLA campus on the north and the west, and by residential neighborhoods with a parking permit districts on the south and east.[1] The average cruising speed was 8.5 miles (13.6km) per hour and the average distance driven while cruising for a free parking space in Westwood was half a mile (313m). Added across all cruising drivers over the year, totals 945,000 extra miles (1,520,830.08km) traveled, using 47,000 gallons of gasoline and producing 728 tons of CO_2. On the Vexpan Parking Convention, 2018, Breuner highlighted another dangerous situation for our health. Cruising of cars leads to deterioration of the air quality, because of wear of tires, and loosening rubber particles that can be inhaled. This topic is one researcher are now interested in.

To restrict the search traffic, various apps have been developed. In 2012 and 2013, Leiden Marketing, in collaboration with Centrum Management and VAG/Parking Management, developed an app that not only provides information about the nearest parking place at the destination, but also about the number of free spaces at the larger parking locations (*Leiden, 2014, p22*). There are also apps designed for the online reservation of parking places (Yellowbrick BV, Parking in Rotterdam, Q-park) and apps that can be used while traveling to locate parking places (Driveguide Terberg Leasing B.V).

Several studies have been done to find a model to predict the occupation distribution of the rate of a parking place. Research in Berlin (2015) shows that data mining techniques using the neural gas algorithm and unsupervised clustering in combination with the original temporal relations of the raw data might lead to good prediction results (*Tiedemann et. al., 2015*). Vlahogianni et. al. (2015) studied the short-term parking occupancy prediction in selected regions of an urban road network using neural network models. The models used captured the temporal evolution of the parking occupancy and may accurately predict the occupancy up to half an hour ahead using one-minute data. In both studies data mining techniques were used. These researches show that a method or algorithm can be found to predict the distribution of the occupancy rate for a parking place with short-term parkers.

Although permit holders have a fixed pattern of parking spaces, that pattern is still subject to chance due to unforeseen events. For example, due to the weather, a permit holder could choose to go to his office by car, leaving an extra parking space empty. Furthermore, parking is also influenced by other factors such as the day of the week and the time of the day. It may be that there are fixed market days in the week attracting different visitors (*Tiedemann, 2015*). And there may also be holiday months in which not only permit holders but also others more often choose to use the car. Research should therefore ensure that the indication of the number of empty parking spaces in a neighborhood is reliable for any type of weather or the time of the year.

Several studies have been done to find a model to predict the occupation distribution of the rate of a parking place. In this report three are mentioned. In the project "Parking Management and Modeling of Car Park Patron Behavior in Underground Facilities", Caicedo et al *(2006, p1)* investigated the behavior of parking patrons in underground parking facilities, a common type of facility in Barcelona, Spain. To model patron behavior, commonly known desegregated models based on the random utility theory were adapted to

[1] See for more background information about this study the book: The High Cost of Free Parking's (Donald Shoup)

facilitate an understanding of how parking patrons decide to use a particular garage level and determine their preferences for a particular garage level. The decisions made depend on the accuracy and the convenience of the information offered. The study finds that an intelligent parking management system that tells a customer the exact locations of the available spaces is of great benefit to patrons and in the long run is a cost-effective alternative to operators.

A research project entitled "Concept of a Data Thread Based Parking Space Occupancy Prediction in a Berlin Pilot Region" was done to develop a prediction for an estimated occupancy of the parking spaces in the pilot region for a given date and time in the future. For this project the data was collected online by roadside parking sensors developed within the project. This research was mostly done with data mining techniques. As it is assumed that the reason for a change in the parking behavior depends on hidden variables, an unsupervised clustering method is used to identify the best matching class. Hereto the neural gas algorithm is used. Then based on these results a prediction model is composed. The combination of a machine learning clustering method and the original temporal relations of the raw data was supposed to lead to good prediction results in reality *(Tiedemann et., 2015).*

The study "A Real-Time Parking Prediction System for Smart Cities" conducted by Vlahogianni et. al. (*2015*), exploited statistical and computational intelligence methods for developing a methodology that can be used for multiple steps ahead on-street parking availability prediction in "smart" urban areas. This model takes real-time parking data, obtained by an extended parking sensor network available in the "smart" city of Santander, Spain. They introduced neural networks for the prediction of the time series of parking occupancy in different regions of an urban network, distribution. The neural networks adequately captured the temporal evolution of parking occupancy and may accurately predict occupancy up to half an hour ahead by exploiting one-minute data. A set back of this study is that the proposed approach is tested on limited data that may not claim to be representative of the monthly variations in parking demand. Moreover, a critical limitation of the present approach is the lack of traffic data that would have provided a more consistent formulation of the parking prediction problem to the evolution of traffic demand.

In this study a mathematical model will be composed using basically mathematical concepts. For known data, the initial distribution of the number of cars on time t, is a canonical vector with one non-zero entries equal to one. If number of cars equals j, then the $j+1$ entry equals one; after all the probability for being is that state is one *(Liu, 2010 p163).* Since the number of cars is binomially distributed, for unknown data the initial distribution is estimated with the mean fraction of cars at time t.

The n-step transition probability matrices are found with the probability distribution of the transfers. The transfer variable is found with the differenced series of the number of cars or the net added number of cars at time t. ($Z(t)=N(t)-N(t-1)$). Another way to determine the transfer variable is to define the net added number of cars as the difference between the number of arrivals and the number of departures ($Z(t)=A(t)-D(t)$).

For the actual predictions, three basic Markov chain models are used: First order Markov chain model *(Ross, 2010).* Higher order Markov chain prediction model as described by Ching, Ng and Wai *(Ching et. al., 2006).* Higher order Markov chain model with triples. This is a model that has a combines three states in one and uses one step transitions.

The idea of taking an extra lag/factor/point into account originates from Raftery *(Raftery, 1985).* That model was extended to a more general higher order Markov chain model that takes the influence of different lags into account (*Ching, 2006, p113*). Higher order Markov chain models do assume that the current state depends on the last k states and are especially useful when an evolution of a series tends to be non-linear *(Ching et al, 2013, pp. 141)*. The mathematical validation for this model is extensively explained by *Wai-Ki et al (2006, chapter 6); Ching et al (2008),* and *Liu Tie (2010).*

The normed squared column sampling techniques of random numerical linear algebra explains how to find a so-called “random sketch” from the original matrix. It is assumed that this sketch has the same properties as the original matrix *(Smetana dr. K., 2018, page 61-69).* In simulated annealing non-homogeneous Markov chains can be partitioned in homogeneous Markov chains (*Hurink, 2017, Lecture 6, p15).* Homogeneous Markov chains are time independent and just see two time points: a start time and an end time; the intermediate time points do not influence the transfers. Non-homogeneous Markov chains are time dependent and associates each transfer with a time point between the start time and the end time of a set of transfers *(BachMaier S, 2016).*

.

1. Problem Description and the Research Variables

This section contains the problem description, the research topic and sub research questions followed by a brief discussion of the research variables. In this report parking space refers to a parking area designed for one single car and a parking place refers to the set of parking spaces.

1.1 Problem Description and Research Questions

In busy cities like Amsterdam finding a parking place is a problem. To reduce cruising traffic ARS Traffic & Transport Technology (ARS T & TT) wants to develop software to inform drivers of the number of empty parking spaces in a nearby neighborhood. They want to have more knowledge and insight in the actual parking distribution of the rate.

Once or twice a day a scan-vehicle passes in the whole neighborhood to scan the vehicles. So, there is some scan data that gives insight in the distribution of the occupancy rate of the past. At the parking place there are two significant types of parkers: 1) the long-term parkers, most of the times the permit-holders, and 2) the short-term parker, most of the time the visitors. Both types of parkers have different parking behaviors. Using the typical characteristics of the parking behavior the company simulated the situation at a large parking place in a neighborhood. This simulation is done for a smaller part of the parking place just as sensors would have done that. The company wants to have a mathematical prediction model for the distribution of the occupancy rate in a neighborhood based on the evolution of the number of short- and long-term parkers as conveyed in the data base of the “sensored” part of the parking place. Such a model should be able to use the available simulated data and the scan data to predict the number of cars at the parking place after a number of minutes.

1.2 The Research Topic

The company wants to know to what extent predictions could be done for the parking occupancy in a neighborhood based on data available to ARS T&TT. Hence, the main research topic for an OR student would be *to find a Markov chain-based prediction model for the distribution of the occupancy rate of the parking place in a neighborhood.* In order to find this model, the following sub questions are considered: What is the optimum fraction of parking spaces that should be equipped with a sensor? What fraction of the parking place should be equipped with sensors? How important is knowledge about the distribution of parking times for visitors and for permit holders? What is the sensitivity of the number of scans per day and the distribution of the scans over the day? Are there other data sources that can provide extra information?

The answer on the first sub question could help one to determine if the data-set is well chosen. It could also help to estimate the a priori error and thus to determine a tolerance range for the a-posteriori error. The expectation is that these errors help to adjust the performance of the model. Generally, knowledge about the distribution of a variable gives a better picture of the location measures such as the mean and the expected value. Moreover, it reveals if the distribution is a joint distribution that should be split. Knowing how many scans are needed each day and at what time period they should be taken can help to find a data set that more adequate represents the detailed situation as generated by the sensor, and in this way even exclude a huge investment in sensors. An answer on the last sub question will only lead to a better model, maybe even a simpler model.

In this report PARK200 refers to the simulated or the “with sensors equipped part of the parking place: (200 parking spaces) and the term PARK1000 implies the whole parking place consisting of 1000 parking spaces.

1.3 Research Variables

The research variables in this study are the type of parker, the day, the time, the number of arrivals, the number of departures, the numbers of cars at the parking place and the net added number of cars at the parking place. These variables are deduced from the data set that describes the situation on PARK200.

1.3.1 Type of Parker Based on Parking Time

The users of this parking place are split into two groups: Long-term and short-term parkers. As it cannot be seen from the data set whether a parker is a permit holder or not, the parking time will be used to identify these two groups. The parking time or parking duration is the total number of consecutive minutes in which a vehicle is parked in the neighborhood. The time starts running from the moment a car is registered as an arrival in a parking space until the next point in time in the system that the same parking space is empty. It is assumed that the parking time is an integer value running from one to 1440. The parking time of a car that stays overnight at the parking place is at least 720 minutes.

In this process a user enters the parking place, and if there is a parking space available the driver chooses to stay for a time period in that space, and after that time period he can choose to stay a next period or leave. This approach the process allows one to identify permit holders that come and go a couple of times in the parking place as a short-term parker and visitors who lengthen there stay a couple of times consecutively occupying the parking space as long-term parkers.

An analysis of the parking time helps to determine to what type of user a car at the parking place belongs. The central tendency of a data set is mostly described using the mean, the median and the mode. The mean of the parking time of all parkers who ever visited the parking place according to the given data is 631 minutes while the median equals 216 minutes. This would imply the existence of two groups of parkers with parking times concentrated around these two values. But, only 24% percent of the parking times are between 180 and 650 minutes. Hence, knowledge of the distribution of the parking time is necessary.

Zooming in on the distribution of the parking times gives a better picture of the data sets. To understand the importance of knowledge about the distribution of parking times, one should first understand the definition of distribution. Rumsey (2018) describes the distribution of the parking times as a list or function showing all the possible values or intervals of the data and how often they occur. One way to visualize the distribution is to use intervals for this continuous random variable and draw a histogram. Using granularity and the relative frequency result in the probability density function. The area under the curve in any given interval tells what percentage of the data falls into the interval.

The parking time is bimodal. This is also clear from figure 1.3.1a. The distribution function of the parking time is bimodal, indicating that the process consists of two underlying distributions. These two distributions appear to be centered around 89 (1.5hours) and 812 minutes(13.5hours).

Figure 1.3.1a: Density parking time parkers

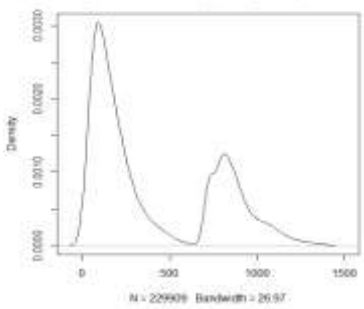


Figure 1.3.1b: Density parking time short-term parkers

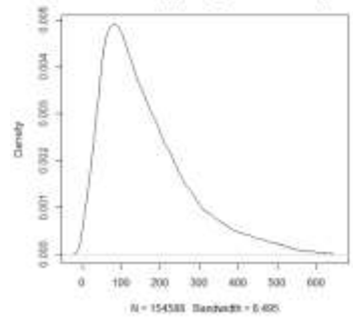
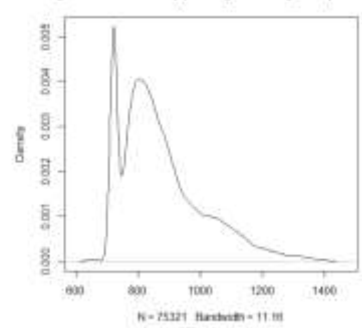


Figure 1.3.1c: Density parking time long-term parkers



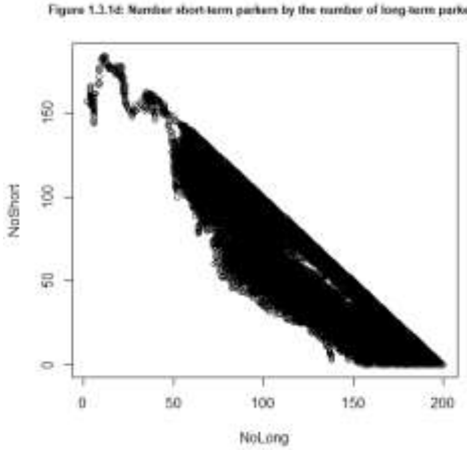
And using the point in between, a long-term parker can be defined as a parker with a parking time of more than 630 minutes and a short-term park is a parker with a parking time of 630 minutes and less. The mean for the short-term parkers equals 162.62 minutes and that of the long-term parkers 873.55 minutes. The nonparametric one-sample Kolmogorov-Smirnov test does not find enough evidence in the data to conclude that the distribution of the parking time is equal to the exponential distribution. Hence the hypothesis that the parking time has an exponential distribution is rejected. See for more details Table 1.3.1.

Table 1.3.1: Summary statistics parking time short- and long-term parkers

Type	mean	SD	median	mad	max	range	skew	kurtosis
Short	162.62	110.55	136	94.89	629	628	1.18	1.27
Long	873.55	132.76	842	109.71	1438	808	1.12	1.01

Generally, it can be said that the shorter the parking time, the more parking spaces available the next minute, something that is welcome especially when the parking demand is high. As the parking place is limited, it is expected on an arbitrary time point *t*, that the number of occupied parking spaces by long-term parkers determines the *maximum* number of cars that *can* enter the system. The number of long-term parkers itself does not necessarily influence the number of short-term parkers.

Figure 1.3.1d: Number short-term parkers by the number of long-term parkers



This is also evident in the banded shaped form of figure 1.3.1d. The number of each type of parkers depends on the demand in each one of the groups and the available space in the parking place. The mean fraction of parkers that belong to the group of short-term parkers equals 0.4618.

A 95% confidence interval for the fraction of short-term parkers within the group of parkers is [0.4602, 0.4634] and for long-term parkers [0.5355, 0.5398]. From here it is clear that the fraction of long-term parkers is on average more than the fraction of short-term parkers on the parking place.

1.3.2 The Day of the Week Based on Date

The day of the week implies one of the 7 days in the week that an observation is done. This variable is deduced from the date. The date is the number of the day on which an observation is done, whether a parking space is empty or occupied. The number of these 500 dates ranges from 0 to 499. This number indicates the number of days that have elapsed since the first observation. Applying the modulo 7 operator+1 on the date results in the numbers 1, 2, 3, 4, 5, 6 and 7, each of which can be associated with a day in the week.

The data set contains for day 1, day 2 and day 3 each 51,840 observations and for the rest of the days each 51,120 observations. Regular activities on a specific weekday in the area of the parking place, could influence the demand of parking. Tsestos et. al. (2015) have shown that distribution of the occupancy rate of the weekday do differ from that of a weekend day. A study in a Berlin pilot region relates in 2015 that the occupancy rate differs also for weekdays. In the plot here below the distribution of the number of short-term parkers and long-term parkers reveal that there are some differences especially for the 7-th day.

Figure 1.3.2a: Distribution number of short-term parkers

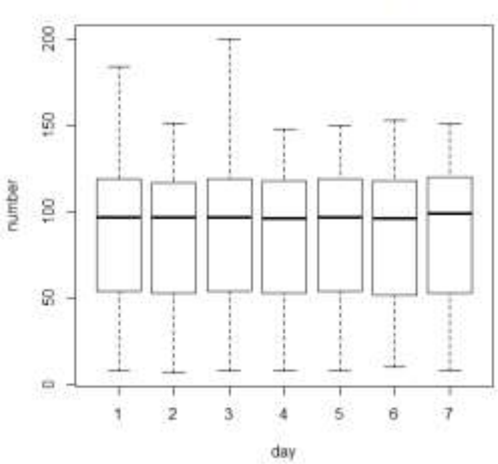
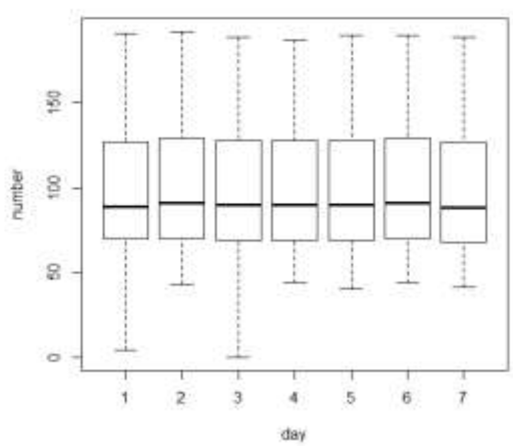


Figure 1.3.2b: Distribution number of long-term parkers



The boxplots show that there are both similarities and differences in days. Therefore, days will not be clustered in this study; the data for each day will be kept separate.

1.3.3 The Time of the Observation

Measurements are done between 9.00h and 21.00h: The time of the observation or briefly the time is the minute of the day on which an observation is done whether a parking space is empty or occupied. The time is indicated in whole units of one minute and runs from 0 to 719. If time equals for example 61 then the actual time is 10.01h. Based on the law of strong numbers, the number of cars is aggregated by time within each group such that patterns in the temporal evolution of the number of cars can be made visible. See figure 1.3.3.

It is clear that depending on different “linear” patterns of the graph of the short-term parkers a day should be divided in more than two time periods; for long-term parkers two periods would be sufficient. Obviously one can use the next time periods: 0-30, 30-179, 179-218, 218-313, 313-420, 420-500, 500-719 to evaluate the process.

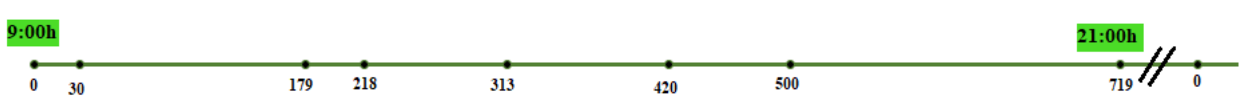


Figure 1.3.3a: Average numbere of short-term parkers on PARK200 by tim

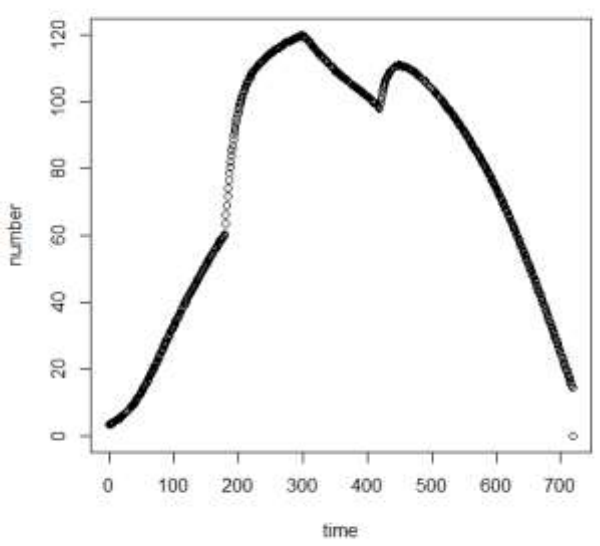
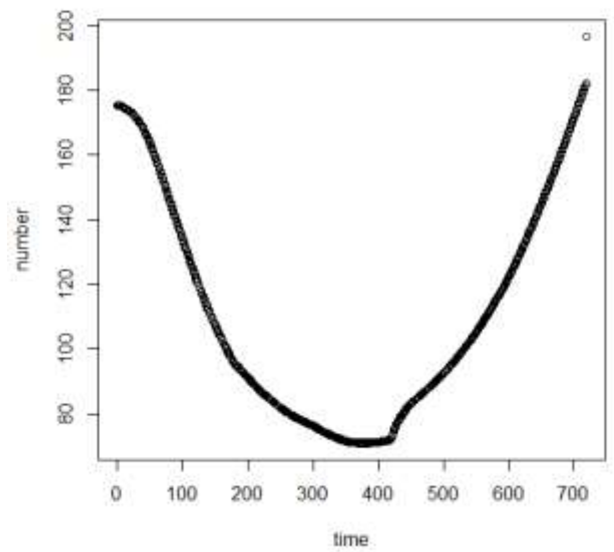


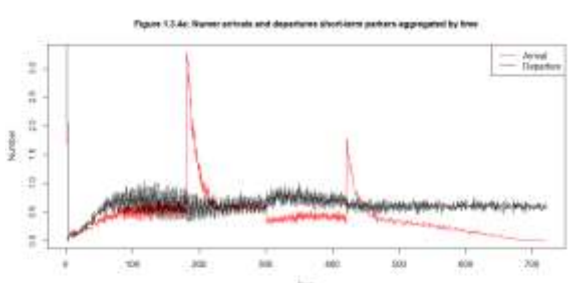
Figure 1.3.3b: Average numbere of long-term parkers on PARK200 by tim

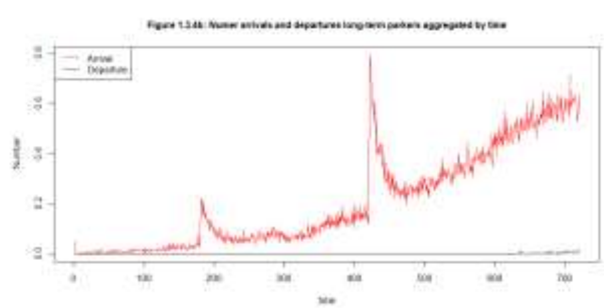


However, one can also choose another partitioning of the day (See section 3.3). Testing the hypothesis of no correlation with the Kendall correlation test shows that at significance of 5% one can conclude that the data does not proof that one can deny correlation between the time and the number of cars at the parking place (all p-values are zero).

1.3.4 The Number of Arrivals and Departures

An arrival or new parker that enters the system and choses to occupy an empty space, will be counted as an arrival. Each car that is in the system and chooses to leave the parking place is seen as a departure. At significance level of 5% a Chi-square test applied on series of the number of arrivals of the short-term parkers and the long-term parkers, shows that the data does not provide enough evidence to conclude that the series are not Poisson distributed. Hence it can be stated that the inter arrival times are exponentially distributed and thus memoryless. With a significance of 5%, a t-test for paired observation reveals that it cannot be said that the mean number of arrivals of the short-term parker does not differs from that of the long-term parkers. Sixty six percent of the arrivals are short-term parkers with an overall mean arrival rate of 0.42 per minute and 34% are long-term parkers with an overall arrival rate of 0.2 per minute. These figures do not take the influence of the day and time into account.





The departure process in both groups can be modeled as a binomial process. For each group, there is a number of cars at the parking place and with a probability depending on time *t* a car within in a specific group departs. As a car driver chooses with a fixed probability if (s)he would be a long-or short-term parker it can be expected that the long-term parkers and short-term parkers do have their own departure pattern and therefore the probability of departure is time dependent. What gives more insight into the need to take different time periods into account is the number of arrivals and departure of both groups. The different behavior patterns for both the short-term parkers and the long-term parkers do justify the idea to split the day into several time periods. See also figure 1.3.4

1.3.5 The Net Added Number of Cars

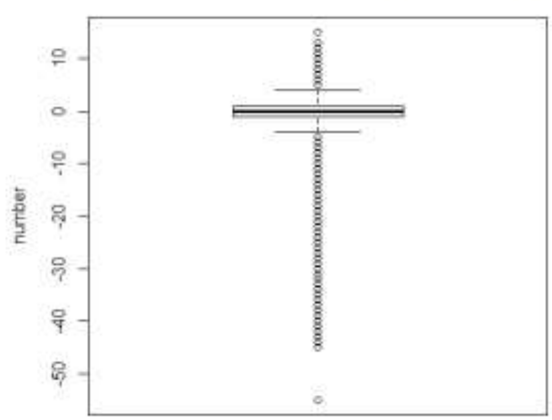
The net added number of cars is in fact the difference between the numbers of cars on two consecutive time units. It represents the real change in car for the next minute after arrivals and departures. The maximum net added number of cars equals 15, the mean 0.00009 and the minimum -55. The boxplot of this change in cars gives a picture of the distribution of the net added number of cars. The impression is that there are quite some observations with an extreme net added number of cars. Interesting is that net added number of cars less than -8 consists can be found most of the time at the beginning of the day (time=0, see Table 1.3.5).

Table 1.3.5: Number of observations in interval of net added number of cars

Net added number	Total Number	Number at time=0
⟨←, -5]	527	417
⟨←, -6]	413	402
⟨←, -7]	388	387
⟨←, -8]	387	377

This reveals that there is an interesting time period that should be isolated; the so-called night time period or the period from 719 till 0 minutes. For time=0 there are 500 observations measured.

Figure 1.3.5a: Distribution Net added number of cars



1.3.6 The Occupancy Rate Based on the Number of Cars per Minute

In this study, the dependent variable or the predicted variable is the occupancy rate. The occupancy rate is the fraction of the spaces at the parking place that are occupied at a specific time point *t*. This rate is expressed as the next fraction:

$$\frac{(\text{total number of occupied spaces at time } t)}{(\text{total number of parking spaces})} \qquad (1.3.6)$$

In this study the focus will be on predicting first the number of cars in the parking place at a certain time point, and then using the relationship in (1.3.6) to compute the occupancy rate. The number of cars at time *t* is a variable that is derived from the set of occupation indicators of all parking spaces at that time point.

The occupation indicator is a dummy variable that indicates whether a parking space is occupied or not. Indicator zero indicates that the parking place is free, and one means it is occupied. Noteworthy is that for a single minute a parking space can only have one occupancy indicator. Hence, the number of occupied places equals the sum of all occupation indicators at that time point.

The distribution of the number of cars is bimodal, resulting mainly from the joint distribution of short-and the long-term parkers. The occupation rate of the number of cars is centered around 0.89 and 0.98. The mean fraction of spaces that are occupied by respectively the short- and the long-term parkers is 0.38 and 0.55. The minimum numbers of cars at the parking place is 128 and yielding a minimum occupancy rate of 0.64. See for more details about each one of the groups Table 1.3.6.

Table 1.3.6: Statistics of the number short and long-term parkers per minute on PARK200

	No Cars			No Short			No Long		
	No	Arrival	Dep	No	Arrival	Dep	No	Arrival	Dep
Minimum	128	0	0	0	0	0	2	0	0
Mean	185.78	0.64	0.64	75.61	0.43	0.64	110.17	0.21	0.0004
Maximum	200	159	58	184	157	58	200	6	1

In table 1.3.6 the overall mean of the number of large term parkers is more than that of the short-term parkers. A maximum of 200 long-term parkers was observed at the parking place. In the group of the parkers the mean fraction of 60% exist of long-term parkers. In the remainder of this study relevant statistics will be linked to the time period and the day of an observation.

2. The Available Data Set

ARS T & TT made two files available for this research namely 'Sanfiles.csv' and 'Occupation.csv'. The first file made available by ARS contains data collected over a period of two years by the scan vehicles. The file with scan data ('Sanfiles.csv') contains for each line the date, the time the scan vehicle started, and in each one of the 1000 columns one indicator for the occupation of a parking space.

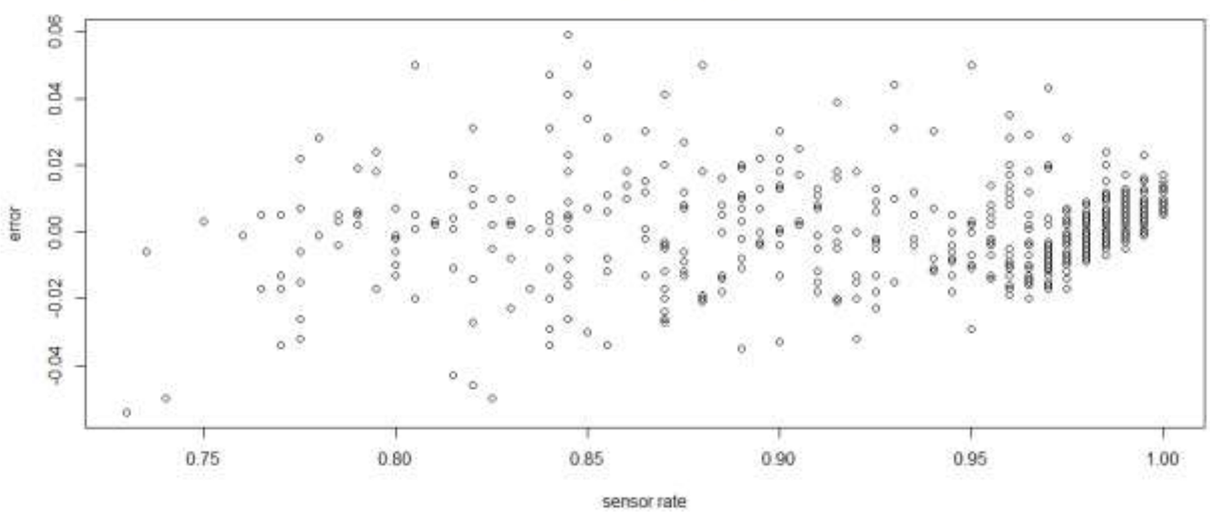
The rest of the data contained in the occupation file is collected by the company by simulating the situation in a neighborhood partially. This file indicates for the first 200 parking spaces of PARK1000 for each line the day, the time and the occupation indicators as if registered by “sensors. When generating data for that fictive neighborhood, ARS uses actual data from Amsterdam and includes the parking behavior and the ratio between the number of visitors and permit holders. It contains 360,000 simulated observations for a period of 71 weeks and 3 days.

The company wants the file with the “sensor” data, to be used to train the model proposed in this report. Hence, the values for the research variables as named in section 1.3 are deduced from this file. The Scan-data files should be used to test or validate the algorithm. In the remainder of the report the simulated data will be referred to as the sensor data or the data coming from PARK200. The sensor rate is then the rate resulting from the sensor data. The data coming from PARK1000 a result from scanning the neighborhood, will be referred to as scan data.

2.1 The A-Priori Error

The time registered for the scan data is in fact the starting time of the scan procedure and does not corresponds with the time registered by the sensors. So, it makes sense to focus on the a-priori error or the difference in the occupation rate of PARK200 (sensor data) and PARK1000 (scan data). This error varies from -0.054000 to 0.059. The mean error is 0.001035, the median equals 0.001, 25% of the data has an error less than -0.007 and 25% has an error of more than 0.008. Moreover, 5% of the differences are less than -0.020 and 5% exceed 0.023. Based on the latter a mean range of 0.043 will be used as an ideal upper bound for the mean range of the a-posteriori error.

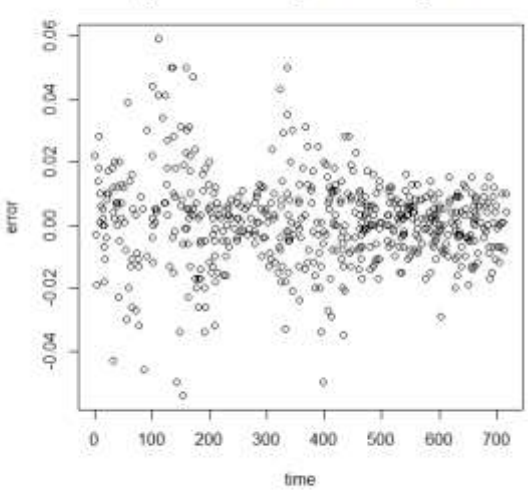
Figure 2.1a: A priori error by sensor rate



A closer look at the scatterplot of the occupation distribution of the rate on PARK200 and the a-priori error reveals some regularities. When the sensor rate is high the variability in the difference in measure is not too high. There seems to be a periodic relation between the sensor rate and the a-priori error. In the remainder of this study, it will be checked whether this relation could be detected and used to predict the number of cars on PARK1000 deduced from the scan data.

The magnitude of the a-priori error could also be related to the time period or the time of the day. Understandably during busy periods when the parking place is full, the difference will not be too large. On the other side, large a-priori errors can be found in busy periods with a low occupation rate. As it is not the aim in this report to correct the error made in the measurement, the mean range for the a-priori error will be used as a bound for the mean range for the predicted interval. The information of PARK200 is used as input for the model as it corresponds to the its scan information.

Figure 2.1b: The a priori error by time



So, the error the scan vehicle makes is not included in the model, but the prediction is adjusted afterward such that the rate resulting from the scan data could be approximated better. Hereto it is assumed that the error is normally distributed on each one of the eight part of the days. It is also thought that all data points from the scan file are needed to adjust the prediction result. See for more details section 2.2.2.

2.2 The Number of Parking Spaces with a Sensor

In this section the sensitivity of the fraction of sensor-equipped parking spaces will be considered. This will be done using two approaches. The first approach states that the number of 200 spaces is representative for the parking place of 1000 spaces. The second approach is that the minimum number of parking spaces that should be equipped is unknown.

2.2.1 The Minimum Number of Parking Spaces with a Sensor

As PARK200 represents the total parking place it is assumed that, *p(t)*, the fraction of occupied spaces at time *t* at PARK200 is a good estimator for the similar fraction on PARK1000 at time *t*. The occupation rate on PARK1000 should be in a 95% prediction interval of predictions done with a model based on data from PARK200.

Suppose Y is the number of occupied parking spaces at PARK1000. Then $Y \sim bin(n, p(t))$, where *m* equals the total number of parking places(*n*=1000) and *p(t)* the fraction of occupied parking spaces at time *t* (*t*=0, 1, . . . , 719). The variable X represents the number of occupied parking spaces at PARK200, with $X \sim bin(m, p(t))$ (*m*=200). Here it is he hypothesis is that the 'success rate' for both distributions are equal. The borders of the prediction interval for each time point would be:

$$\hat{p} \pm t_{\frac{1}{2}\alpha,(n-k+1)} \sqrt{\frac{\hat{p}(t)*(1-\hat{p}(t))}{m^2} + \frac{\hat{p}(t)*(1-\hat{p}(t))}{n}}, \qquad (2.2.1a)$$

where $\hat{p}$ is an estimator for the mean fraction of occupied spaces at PARK200 and k=2, the number of involved random variables (See appendix 1 for more details). From the 598 scanned moments 93.3% of the deduced rates are in the associated 95% prediction interval, with mean range 0.063. For a 90% prediction interval the percentage is 91.5%, and the mean range 0.053. This implies that if the ideal mean range of 0.043 cannot be found, a mean range of 0.063 could also be tolerated.

A theoretical percentage of 10.20% of the posteriori errors lies within the range deduced from the a-priori error into account ([-0.020, 0.023]). Under the null hypothesis that 90% of the predicted errors lies somewhere between -0.020 to 0.023 it can be concluded that at significance level 10% the data set does not provide enough proof to believe that the rates from PARK1000 lies in a prediction interval the with range 0.0043. So, with this data set it cannot be expected theoretically to find a model that predicts the rate on PARK1000 without tolerating a posteriori error larger than the a priori error.

Hence it can be concluded that a model based on the information given by 200 spaces is not able to predict the occupation rate at PARK1000. Apparently, more spaces should be equipped with sensors. More sensors do have another benefit for a prediction model. The more sensors in the parking place, the larger the value for *n*, the smaller the range of the prediction interval (2.2.1) the more accurate the predicted rate resulting from a good model.

Depending on the error the company wants to make the number of parking spaces (*m)* can be estimated using the ranges in 2.2.1. Suppose the absolute error the company wants to make equals error ε. Then the difference of the range and the mean is error such that:

$$t_{\frac{1}{2}\alpha,(n-k+1)}\sqrt{\frac{\hat{p}(t)*(1-\hat{p}(t))}{m^2}+\frac{\hat{p}(t)*(1-\hat{p}(t))}{n}}=\varepsilon \qquad 2.2.1b$$

Under the assumptions that $\alpha = 5\%$, *n=200*, $\hat{p}(t) = p$, ε is well chosen by the company, the m can be found by solving the equation in statement 2.2.1b. It should be remarked that the fraction $\hat{p}(t)$, is time dependent. So, by deciding what error the company wants to make it should be taken that also into account.

2.2.2 A Lower Bound for the Maximum Number of Parking Spaces with a Sensor

Another approach to determine the sensitivity of the fraction of spaces that is equipped with sensors assumes that the parking place consists of maximum 1000 spaces. Equipping the whole population of parking spaces with sensors would be the best but most expensive choice. Can the company suffice with equipping fewer parking spaces? In this section the concept of random numerical linear algebra (RandNLA) and the concept of the rank of a matrix will be used.

One main concept of RandNLA algorithms is constructing a so-called random sketch of the considered matrix by random sampling and then using the sketch as a surrogate for the computations (*Smetana, 2018, section 8.2).* Matrices are seen as linear operators, such that the role of rows and columns become more central. Based on the knowledge of the space a fixed number of columns can be chosen according to the simplest non-uniform distribution known as l_2 sampling or norm-squared sampling, in which p_i is proportional to the square of the Euclidean norm of the i^{th} column:

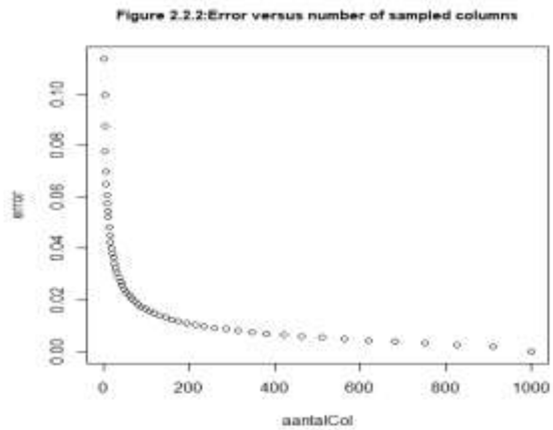
$$p_i = \frac{\|A_i\|_2^2}{\sum_{i=1}^{n}\|A_i\|_2^2} \qquad (2.1.1)$$

Each time a lower number of columns is sampled. This implies that the number of parking spaces is gradually reduced. The occupation rate of this random sketch or reduced parking place is computed and its mean deviation from the universe. It is expected that the mean error would be zero on the long run.

The occupation rate on the whole parking place depends on what happens each minute in each parking space. Hence the parking place can be seen as an *m* x *n*-dimensional space where *m* is the number of the observation and *n* the number of the parking space. The space dimension is represented by *n* and the time dimension by *m*. The data file that represents the situation at the parking place is a matrix full of indicator variables. A row shows for a single time for every parking space whether it is occupied. A column shows for one parking space for every minute whether it is occupied. In this way the parking place is set-up by 1000 parking place vectors.

Random column sampling is simulated 100 times. In each simulation the number of columns (parking spaces) is gradually decreased by the factor $\frac{10}{11}$. For each *n* columns (parking space) that are sampled, the occupation rate is computed for both the sampled parking space and for PARK1000 together with its absolute difference or error. In figure 2.2.2 the mean error for the occupation rate aggregated by number of columns is plotted.

There is a negative correlation between the number of columns and the absolute mean error with respect to the universe of 1000 parking spaces; the less column, the larger the error. This relation between the number of columns (y*)* and the mean error (x*)* could be approximated the with *y=392.68*exp(-63.62x).* Depending on the error the company allows it could decide how many spaces should be equipped with a sensor. The number found with the exponential relation is in fact a lower bound for the maximum number of spaces that should be equipped with sensors.



3. Mathematical Description of the Problem

The total number of cars is generated over consecutive minutes of consecutive days. The resulting sequence of the number of cars at the parking place from minute to minute is in fact a time series. The observations in this time series, have a recurring pattern for measurements made on the same day and the same time, such that the occupancy rate is time-dependent. In this report it is the aim to find a mathematical model that takes not only the factor time but also the number of the short-term and long-term parkers into account. Designing such a model for software is a challenge because parking cars itself is a stochastic process or a succession of accidental outcomes.

To make this research operational, the scope of the study is restricted, and assumptions are made to model the real-life situation. First the restrictions and assumptions are discussed in this section.

The next section shows why and how this problem can be modelled as a discrete time inhomogeneous Markov chain. “Markov chains” is the core business of operational research. Therefore, it was decided in this research to predict the occupancy rate, with Markov chain prediction models. Admittedly, it should be remarked that research has shown that Markov chain prediction models lack accuracy when history matters *(Wu T., Gleich D., 2017 p.1*). Nevertheless, it is expected that there are enough mathematical “tools” to correct prediction flaws. As the use of Markov chains will be eminent the focus will be now on predicting the probability distribution of the number of cars at time *t.* The third section proceeds with a discussion of the time series properties that are included in this research. The last concludes with a mathematical description of the problem.

3.1 Restriction and Assumptions Needed to Model the Problem

The first restriction in this research is that the parking problem only addresses on-street parking in a closed neighborhood, or a neighborhood with a fixed number of parking spaces. This study is based on data coming from simulating the parking process at a part of the whole parking place. The simulation is based on the way sensors work in the parking garage of Schiphol airport. This study does not test how well the sensors do reflect reality. Nor is this research intended to ascertain whether the provided data had been correctly simulated. In addition, it will not be examined whether it is correct to take the starting time of scanning the neighborhood as the actual time for scanning an arbitrary parking space in the parking place.

The number of free spaces at the parking place depends largely on the parking time of the users. The shorter the parking time the more parking spaces available each minute. Hence this research assumes that predicting the number of free spaces requires acknowledging the parking time and consequently the existence of two distinct groups in the system: the long-term parkers and the short-term parkers. Both visitors and license holders can choose which open parking place they will use. Every parker is allowed to use a parking space for the period of time determined by him or her, provided the payment is made.

It is assumed that the users of the parking place act independent from each other. It is thought that a driver indicates upon arrival how long (s)he will stay in the system and customers do take decisions independent from each other. In this way each driver that comes into the system decides for himself if he is a short-term parker or a long-term parker. The number of short-term parkers and the number of the long-term parkers that are added each minute to the system depend on the available space at the parking place, the parking demand and the parking behavior. The inference for each minute is therefore that the number of short-term parkers does not depend on the number of long-term parkers in the parking place. It is also assumed that the situation on PARK200 represents the situation on PARK1000.

As it is not clear from the data what happens exactly between 21:00h and 9:00h it is assumed that all cars present at 21.00h do stay overnight. Hence, short-term parkers do not stay overnight at the parking place and every car present at the parking place at the end of the day (minute 719) is a long-term parker. This assumption tends to correspond with the reality. Generally, a neighborhood is not used for business activities in the evening. During the day its parking place becomes an extension of the city. As most of the shops and business places in the city are closed, it is expected that there are enough free parking spaces in the city; there is no extension of the city needed in the evening hours. It could be that short-term parkers occasionally use the parking place in the evening. The first time the data is recorded it is not known how much of the cars did stay overnight, so all the cars at that moment are treated as arrivals.

3.2 A Non-Homogeneous Discrete Time Markov Chain

Cars do enter the parking place to look for a free parking space. If they found one, they stay for a while; in case no parking place is found they 'leave'. As it is not clear how to understand 'leave', it is simply used in this study to bring over the idea that the number of cars never exceeds its maximum. The situation at the parking place is in fact a M/G/c/c process. This is a process where: 1) the arrival times are exponentially distributed and therefore memoryless; 2) the parking time of an individual car represents the service time and follows a general distribution; 3) the parking place counts *c* servers or parking spaces and no waiting places, such that 4) the maximum number of cars that can be parked in the system at one time point equals, *c,* the number of servers.

Every customer who enters the system is supposed to select a free server randomly. The probability that an arrival occurs in a certain unit of the parking place, is by assumption equal for all other units of the parking place. Moreover, it is assumed that the number of arrivals that occur in an arbitrary unit of the parking place is independent of the number of arrivals in other units. Hence it can be assumed that the Poisson properties for the arrival rates do hold also on PARK200. Transitions do take place each minute. The transition from the evening to the morning that means form minute 719 to minute 0 is considered to be a transition done in one minute or step. Each time period the occupation of PARK200 is registered, resulting in a dynamic sequence of the number of cars is known each time period. Hence, this process can be modelled as a Markov chain.

The Markov chain can be defined basically as a stochastic process $\{X_n , n = 0, 1, 2, . . . \}$, with a finite number of possible values or states, E={0, 1, 2, . . . , 1000} and $X_n = k$, implies that the process is in state *k* at time *n*. Each state represents the number of occupied parking spaces at a certain time point. The one step transition probabilities are time dependent and stored in a matrix

$$P(t)= \{p_{ij}(t), i, j \in \mathrm{E}\}\ p_{ij}(t) \geq 0, \sum_{\mathrm{i,j} \in \mathrm{E}} p_{ij}(t)=1.$$

This Markov chain is not homogeneous because the evolution of the system depends on time. Moreover, its underlying arrival process is a non-homogeneous Poisson process (see graph 1.3.3c) (*Ross, 2010, p372*). The Markov chain has different transferring behavior patterns. Having a transition probability matrix for every time interval of one unit, would imply having a model with hundreds of transition probability matrices. That is not efficient for both the user and computer programs.

In simulated annealing a non-homogeneous Markov chain can be seen as an infinite number of homogeneous Markov chains of finite length each (*Hurink, 2017, Lecture 6, p15*). In this study the "infinite" countable number of homogeneous parts are clustered in different consecutive periods. Each one of his cluster of homogeneous parts is thought to be a homogeneous Markov chains of finite length. As Bach Maier notes that a homogeneous Markov chain sees only two time points: a starting time and an end time. Therefore, the transition probabilities of this non-homogeneous Markov chain will be linked to intervals ranging from start time, t_1, till end time, t_2, such that h=$t_2 - t_1$, is the time span of the interval in which homogeneity is assumed (*Bach Maier S, 2016 p 18*). It is then assumed that for $t\epsilon[t_1, t_2]$, the transition probabilities at time *t*, can be estimated by the transitions of the observations in associated time interval.

There are different ways to define these intervals in which homogeneity is assumed. One way is by using the partitioning of the day *d* in the eight periods deduced in section 1.3.3. Another way is a partitioning of day *d* in consecutive disjoint time intervals of *h* minutes starting from the first minute (*t=0)*. It can also be assumed that the Markov chain is homogeneous in a radius of *s* minutes from the actual time *t*, such that the associated interval equals [*t-s, t+s*]. The transitions in these time intervals will be used to estimate the transition behavior and probabilities at time *t*, on day *d*. Hence in this report a transition probability matrix *P(t)*, is in fact a matrix containing the transition probabilities of a so-called homogeneous part of the non-homogeneous Markov chain associated with a time interval containing time *t*. **From now on, it will be referred to as $P_{\mathcal{J}_t}$for short-term parker and $Q_{\mathcal{J}_t}$for long-term parkers, where $\mathcal{J}_t$ is the time interval containing time *t* and will be used to estimate the transition probability matrix on time *t* at day *d*.** No extra index is needed for the day in this notation; the day(s) connected to time interval $\mathcal{J}_t$ is (are) automatically determined by the actual time of the process and the definition of the time interval $\mathcal{J}_t$.

Moreover, it will be assumed that these homogeneous parts of the Markov chain are irreducible. As it is not clear how to link groups of states from PARK200 to PARK1000, it is chosen to allow transitions for every state on PARK1000.

Further it is assumed that the history, that is the occupation data on PARK200 is known up until the moment of the actual scan when a prediction starts. Consequently, depending on the value of the actual time and day, future values needed to find the transition probability matrix on $\mathcal{J}_t$, the associated homogeneous time interval are unknown. One way to solve this is to assume that information of all similar time periods in the past is needed to determine the transition probabilities associated with the current time interval. Another solution could be to estimate the unknown values by their expectation. A good estimator for the expected value is the mean value of the number of cars aggregated by time and day (See section 4 for the requirements for a good estimator). A t-test for paired observation shows that it cannot be concluded that the difference of the actual values and this aggregated value does not equal zero. For the number of cars, the number of short, and the number of long-term parkers the p-value equals one. Hence these values will be used to find the transition probabilities in $\mathcal{J}_t$. (See also introduction of section 4)

3.3 A Markov Chain with Time Series Properties

For time series it is assumed that the data consists of a systematic pattern and random noise (error). Generally, the systematic pattern of a time series has two components: trend and seasonality.

In this study it will be assumed that the *trend* of the number of cars depends on the transition behavior of the process in a closure around the actual time. In this closure or time interval homogeneity of transferring behavior will be assumed. Beside this the concept of "a forward moving average" will be implemented in making predictions to smooth out the trend and to cancel out large differences at time *t*, *t*=0, 1, 2, …. This will be implemented by choosing a "forward moving interval", [*t*-*h*, *t*+*h*] to determine a "forward moving transition probability matrix".

The next systematic component of the time series is the *seasonal* element. The time series of the number of cars consists of a pattern that is repeated every 720 minutes. The for seasonality adjusted number of cars or the number of cars minus the seasonal component was compared with the number of cars itself. A t-test for paired observations reveals that the difference between the series of the number of cars and the for seasonality adjusted number of cars is negligible (p-value=1). Hence there is no need to split the systematic part of the series and zoom in separately on the trend and the seasonal series.

The *random noise* will be corrected each minute by using a convex combination of the distribution of the predicted value and the distribution of the expected value. The distribution for the expected value at a given day and time will be estimated by the distribution of the average value aggregated by day and time.

3.4 Mathematical Problem Description

In this study the focus will be on the totals per minute on PARK200. The resulting Markov chain of the number of cars at time *t* at the parking place exists of two distinct groups of users who act independently from each other: the short-term parkers and the long-term parkers. Consequently, this Markov chain is in fact a two-dimensional discrete time Markov chain. The data of the *M* parking spaces on PARK200 will be used to predict for time *t*, the distribution of the number of cars on PARK1000 that has a maximum of *N* parking spaces.

On PARK200, $M_s(t)$ and $M_L(t)$ do represent respectively the number of short- and long-term parkers at time t, where $t \in 0,1,\ldots,719$, $M_s(t) \in [\,0, M\,]$ and $M_L(t) \in [\,0, M\,]$. The two-dimensional state $(M_s(t), M_L(t))$ represents the number of short-and long-term parkers at time *t.*

$N(t+w)$ represents the number of the parkers on PARK1000, *w* minutes later, at time *t*+*w* and *k* is the number of historical points that one would like to include in the model (0<$k < t - 1 - t_0$). For the distribution of number of cars on PARK1000 at time (t+w), given the actual state *i*, on PARK200 at time *t*, we want to find $\forall\, j \in E$ the conditional probability:

$$p_{j|i,\, i_{2,}\ldots\, i_k}(t+w) =$$

$$P\{N(t+w) = j \,|\, (M_s(t-1), M_L(t-1)) = (v_1, l_1),\ .\ .\ .\ .\ ,\ (M_s(t-k), M_L(t-k)) = (v_k, l_k)\}$$

, $v_1 + l_1$=*i*, for *i*, v_k, l_k=*0, 1, 2, . . . , N*, *j=0,1,2, . . . , M,* Thus, the distributions from a two-dimensional Markov chain on PARK200 will be used to make predictions the distribution for a one-dimensional chain based on the condition probability stated above. The mathematical concepts needed to solve this mathematical problem can be found in chapter 4.

4. The Mathematical Concepts

This study uses many mathematical concepts based on the assumptions to validate and to sustain the Markov chain models. One frequently used concept is that of the convolution of two independent random integer variables. Independency between groups, allows applying this concept to find the probability distribution of the sum of integer variables. The expected number of cars (L_t), at time *t* is computed with the next formula: E(L_t)= $\sum_{i=0}^{k} x_i^{(t)} i_t$, where *i* is the state or the number of cars in the parking place at time *t* and $x_i^{(t)}$ is the probability of being in that state at that time *t.*

In this research the average often used as a good estimator for the expected value. This concept conveys the idea that the formula to estimate the mean of numbers is "pure" (E($\bar{X}$)=u), efficient (var($\bar{X}$) is minimal) and consistent (as $n \to \infty$, $var(\bar{X}) = \frac{\sigma^2}{n} \to 0$) (*Bolle et. al., 1974, p28-30*). In this report the average of the data will be aggregated by day and time. Moreover, the law of the large numbers can be applied in this data set because of its robustness. This law implies that as the number of observations grows, the influence of a single observation becomes smaller and smaller. Hence, the influence of outliers decreases, and the average of the observations do give a relatively good picture of the expected value of the corresponding population. See the appendix 2 for some more details regarding the convolution and the strong law of the large number.

The discussion in this section focuses mainly on the Markov chain formulas that are used to predict the distribution of the states at time *t.* This section starts with a discussion of the Markov chain prediction models. It addresses concepts for the first and the higher order Markov chain prediction models and concepts needed for the construction of different types of transition probability matrices.

The mathematical concepts quoted in this section regard a one-dimensional homogeneous Markov chains and will therefore be applied in the remainder of the study on parts of the Markov chain where homogeneity is assumed. If an application of a mathematical concept regards the aspect of the non-homogeneity of the chain, it will be mentioned explicitly.

4.1 The First Order Markov Chain Prediction Model

The simplest homogeneous Markov chain prediction model is the first order model. This model assumes dependency on one historical point. The Markov property of the discrete-time Markov chain, that the conditional distribution of X_n given on the states of the past depends on these past states only through the state at the end of time *n* satisfies the following relationship:

$$P(X_n = i_n | X_{n-1} = i_{n-1}, X_{n-2} = i_{n-2}, \dots, X_0 = i_0) = P(X_n = i_n | X_{n-1} = i_{n-1}) \qquad (4.1.1)$$

where X_n and i_n both refer to the state of the Markov chain at time *n (Ross, 2010, p192).*

This relation determines the one-step transition probabilities of the Markov chain from time *(t-1)* to time *t* and can be written as $p_{ij} = P(X_n = i| X_{n-1} = j)$ for *i* and *j* in *E* (*E* is the set of states). Where p_{ij} is the transition probability of going from state *i* to state *j* in one step or one minute. All these transition probabilities are stored in the transition probability matrix, P . More over $0 \leq p_{ij} \leq 1$ $\forall i, j \in E$ and $\sum_{i=1}^{N} p_{ij} = 1, \forall j \in E$, *N* represents the last state in set *E*.

At this point a new matrix is introduced: P_n containing all the transition probabilities of the processes going from state *i* to state *j* in *n* additional transitions. This matrix can be computed from the data but also estimated as the *n*th power of the one step probability matrix, P^n, in homogeneous parts of the Markov chain. The entries of this matrix, p_{ij}^n, do represent the transition probability of transferring from state *i* to state *j* in *n* steps. The probability that, starting in *i*, the process will go to state *j* in *(n+m)* transitions through a path which takes it into state k at the n^{th} transition is represented in the following Chapman-Kolmogorov equations *(Ross et, 2010, p 195).*

$$p_{ij}^{n+m} = \sum_{k \in E} p_{ik}^n p_{kj}^m, \forall n, m \geq 0, \forall i, j \in E \tag{4.1.2}$$

Applying Chapman-Kolmogorov equation in the prediction model allows a prediction model that uses the probability vector at time *n*, to predict the probability vector after *m* steps as follows:

$$x^{(n+m)} = x^{(n)} P_1^{(m)} \tag{4.1.3}$$

For predictions after a minute the most basic form of this Markov chain prediction model reads:

$$x^{(n+1)} = x^{(n)} P_1 \tag{4.1.4}$$

where $x^{(n)} = (x_1^{(n)}, x_2^{(n)}, \dots, x_j^{(n)}, \dots, x_k^{(n)})$ represents the distribution of the states at time *n* or the probability that the process will be in each one of the *k* states at time *n*.

For the non-homogeneous Markov chains the probability matrix is in fact estimated for time *t*, on day *d*, the by using the data from, $\mathcal{J}_t$, the time interval associated with the actual time (See section 3.2 for definition $\mathcal{J}_t$). Taking the non-homogeneity of the chain into account doing prediction with the first order model at time *n* for time *(n+m)* yields:

1) The probability that, starting in *i*, the process will go to state *j* in *(n+m)* transitions through a path which takes it into state *k* at the n^{th} transition is for represented in the following Chapman-Kolmogorov equations

$$p_{\mathcal{J}_t,ij}^{n+m} = \sum_{k \in E} p_{\mathcal{J}_t,ik}^n \, p_{\mathcal{J}_t,kj}^m, \forall n, m \geq 0, \forall i, j \in E \qquad , t \in \mathcal{J}_t, \dots \tag{4.1.5}$$

2) The predicted distribution after *m* minutes starting at time *n*, on day *d*, equals:

$$x^{(n+m)} = x^{(n)} \prod_{t=n}^{m-1} P_{\mathcal{J}_t} \tag{4.1.6}$$

3) For m=1 the basic first order prediction model yields:

$$x^{(n+1)} = x^{(n)} P_{\mathcal{J}_n} \tag{4.1.7}$$

4.2 Higher Order Model with a Combo of States

This higher order model with a combo of states is in fact a special application of the first order Markov chain prediction model. This model is applied on homogeneous Markov chains and considers k historical points as one state. For simplicity k is set to three in this section, such that a combo (combination of states) consists of three consecutive historical time points. In the remainder of this report these combos will be referred to as triples. Under the Markovian property mentioned in t statement *(4.1.1)* a one-step transition probability of being in state i at time n-1 and going to state j at time n is defined as:

$$p_{ij} = P\{(X_{n-2}, X_{n-1}, X_n) = (k, i, j) | (X_{n-3}, X_{n-2}, X_{n-1}) = (m, k, i)\} \tag{4.2.1}$$

The probabilities are derived from the probability distribution list retrieved from the number of all possible events in the associated part of the chain where homogeneity is assumed. The probabilities of all transfers starting with (a, b, c) do sum up to one. $(X_{n-3} = a,\ X_{n-2} = b\ ,\ X_{n-1} = c)$. To reflect the last state as transient state an extra transfer is added to the probability distribution list of the transition probabilities. This extra transfer connects the last event to the last event in the homogeneous part of the chain.

In this context $p_{ij}^{(n+w)}$ denotes the probability of going from state i at time n to state j in w minutes. Starting at time n the predicted probability distribution at time *(n+w)* would be $\forall i, j\ \in E, n, w\ \geq 0$

$$p_{ij}^{(n+w)} = \sum_j \left(p_{i_n j_{n+1}} * p_{i_{n+1} j_{n+2}} * \ldots * p_{i_{n+w-1} j_{n+w}}\right), \tag{4.2.2}$$

where i_n single state at the n^{th} minute. Hence the input for this model is the actual number of cars at three consecutive time points, $X_{n-1}, X_{n-2}, X_{n-3}$ and the transition probabilities in its associated homogenous interval.

The expected value for the number of short-term parkers at time n, is computed using the value of all possible events at time n within the triples and their transition probabilities such that:
$X_n \approx E(X_n) = \sum_j^N j * p_{ij}^{(n)}$, where $p_{ij}^{(n)}$ as defined in statement (4.2.2).

4.3 Higher Order Model for Markov Chains Chin et al

This section starts with a discussion of the higher order Markov chain prediction model a proposed by Chin et. al.. This prediction model assumes that the next state depends on k, a fixed number of consecutive historical time points. Here k is strictly less than the chain length. This model could be interpreted as a special type of the higher order model with a combo of states. While the higher order model with a combo of states links the states by using a joint distribution of the combo of states, the higher order model as proposed by Chin, links the k consecutive states through their distinctive associated probability distribution. For a homogenous Markov chain, the following relationship is satisfied:

$$P(X_n = i_n | X_{n-1} = i_{n-1}, X_{n-2} = i_{n-2}, \ldots, X_0 = i_0) = P(X_n = i_n | X_{n-1} = i_{n-1}, X_{n-2} = i_{n-2}, \ldots\ldots, X_{n-k} = i_{n-k}) \tag{4.3.1}$$

This relation implies that the basic property of Markov chains is extended to the idea that the conditional distribution of the states of the Markov chain given on the states of the past depends on these past states only through the last *k* states at the end of time *n*. The k^{th}-order model for Markov chains assumes that the next state depends on the distribution of the previous *k* states at times *n-1, n-2, . . . n-k* and *k=1, 2, . . , n-1* (*Ching et al, 2008*).

Assuming dependency on the last *k* states the distribution on time *n* could be modeled as a linear combination of the estimated distribution of the series at time *n-1, n-2, . . . , n-k.*

$$x^{(n)} = \sum_{i=1}^{k} \lambda_i \quad P_i \;\; x^{(n-i)} \, , \tag{4.3.2}$$

where:

$x^{(n)}$ is the probability distribution of the states at time *n*

λ_i are real numbers for the parameters of the model in (4.3.2), such that $\sum_{i=1}^{n} \lambda_i \; = 1$ and $\lambda_i \; \geq 0, \; \forall i$

P_i is the *i*-step transition probability matrix

To find the best possible set of parameters for the prediction model in (4.3.2) the stationary distribution for homogeneous part of the Markov chain which will be used. The following theorems and proposition validate the existence of a stationary distribution.

Theorem 2 The Perron-Frobenious theorem *(Smethana, 2018, page 54)*
Let $R = (r_{ij})$ be an element wise non-negative $(r_{ij} \geq 0)$ irreducible N x N matrix. Then:

1. *R has a positive eigenvalue φ which equals the spectral radius of R (φ=$\rho(R)$)*
2. *The eigenvector w, corresponding to φ, can be chosen element wise strictly positive.*

As the probabilities in the rows of the *n*-step transition probability R_n matrices do sum up to one, the first part of the theorem shows that there exist an eigenvalue *φ* equal to one such that the equation $R_n^T\vec{w} = \vec{w}$ is solvable. The second part implies that the solution is in fact an eigenvector with non-negative entries. The next proposition deduced from the Perron-frobenius theorem shows that the eigenvector as mentioned in part 2 of the theorem is in fact the stationary distribution of the irreducible matrix *R* ($\sum_{i=1}^{n} \lambda_i \quad P_i$).

Proposition 1 (Ching et al, 2006b, page 114)
If P_k is irreducible and $\lambda_k > 0$ such that $0 \leq \lambda_i \; \leq 1$ and $\sum_{i=1}^{k} \lambda_i \; = 1$, then the model in (4.3.2) has a stationary distribution $\bar{\boldsymbol{x}}$ when $n \to \infty$ independent of the initial state vectors $x^{(0)}, x^{(1)}, \ldots, x^{(k-1)}$. The stationary distribution is also the unique solution of the following linear system of equations: $\left(\mathrm{I} - \sum_{i=1}^{n} \lambda_i \quad P_i \right)\bar{\boldsymbol{x}} \; = 0$ and $\boldsymbol{1}^{\boldsymbol{T}}\bar{\boldsymbol{x}} \; = 1$. Where I is the m x m identity matrix and $\mathbf{1}$ is an m x 1 vector of ones.

The vector, $\bar{x} \; = (x_1, x_2, \ldots, x_j, \ldots, x_N)$, represents the limiting distribution or the long run proportion of time that the process will be in each state. The time independent vector x is also called the steady-state (stationary) probability vector. An arbitrary entry x_j represents the stationary probability that the system will be in state *j*. In this research the stationary distribution of an irreducible matrix is found with the power iteration of Von Mises *(Smetana, 2018, section 5.2 p.33).*

If the stationary distribution is computed in a proper way with limited data it could be estimated with $\hat{\boldsymbol{x}}$. Hence, it is expected that: $\hat{\boldsymbol{x}} \approx \sum_{i=1}^{k} \lambda_i P_i \hat{\boldsymbol{x}}$. The best model has a minimal error.

$$\text{Error}=\sum_{i=1}^{k} \lambda_i P_i \hat{\boldsymbol{x}} - \hat{\boldsymbol{x}} \qquad (4.3.3)$$

To find an optimal model one should minimize $\left|\sum_{i=1}^{k} \lambda_i P_i \hat{\boldsymbol{x}} - \hat{\boldsymbol{x}}\right|$, the norm of the error. Using the infinity norm would imply minimizing the maximum absolute entry. If the error with the maximum absolute entry is the *l*-th entry the objective function of the optimization problem would be:

$$\min_{(\lambda_1, \lambda_2, \ldots, \lambda_k)} \max_{l} \left| \left[\sum_{i=1}^{k} \lambda_i P_i \hat{\boldsymbol{x}} - \hat{\boldsymbol{x}} \right]_l \right| .$$

This non-linear optimization problem can be transformed into a linear optimization problem. Suppose, $\exists\, \omega \geq 0,$, such that $| \sum_{i=1}^{k} \lambda_i P_i \hat{\boldsymbol{x}} - \hat{\boldsymbol{x}}| = \omega$. Then the optimization problem is transformed into minimization of ω such that $| \sum_{i=1}^{k} \lambda_i P_i \hat{\boldsymbol{x}} - \hat{\boldsymbol{x}}| \leq \omega$ or -$\omega \leq \sum_{i=1}^{k} \lambda_i P_i \hat{\boldsymbol{x}} - \hat{\boldsymbol{x}} \leq \omega$.

The linear optimization problem equals:

$$\min_{(\lambda_1, \lambda_2, \ldots, \lambda_k, \omega)} \omega$$

Subject to

$$\sum_{i=1}^{k} -\lambda_i P_i \hat{\boldsymbol{x}} - \omega \leq -\hat{\boldsymbol{x}}$$

$$\sum_{i=1}^{k} \lambda_i P_i \hat{\boldsymbol{x}} - \omega \leq \hat{\boldsymbol{x}}$$

$$\omega \geq 0, \sum_{i=1}^{k} \lambda_i = 1, \lambda_i \geq 0\ \forall i$$

(Chin et. Al., 2006, Page 117-119, Liu Tie, 2010, p164).

If the optimization problem could not be solved, it will be assumed that all parameters do have the same value. The higher order model can be seen as an extended version of the first order model. It takes more historical time points into account. The distribution is predicted as a linear combination of the distributions of a fixed number of historical time points.

The input of this model is the set of distribution vectors that can be associated with the previous *k* historical points of the process, $\left(x^{(n-1)}, x^{(n-2)}, \ldots . x^{(n-k)}\right)$, together with the set of *n*-step transition probability matrices $P_1, P_2, \ldots \ldots, P_k$. Using the optimal parameters associated with each one of the historical distributions ($\lambda_1, \lambda_2, \ldots \ldots, \lambda_k$), can help one to find a predicted distribution on time n, $x^{(n)}$, that is closest to the stationary distribution.

4.4 Construction of the Transition Probability Matrix P

The data related to time interval $\mathcal{J}_t$ as introduced in the last part of section 3.3 will be used to find the transition probability matrix on time *t*; this includes not only taking the actual time into account but also the day *d*. Focusing on the concept of the expression 'transition probability' restricts the study to finding the probabilities of the net *change* from state *i* to state *j* in one minute. The aim is now to find the probability that in fact *k* cars are added to the current number of cars. Hereto the probability distribution of the net number of cars that are added to the system each minute is computed. The probabilities of this distribution are used to construct the transition probability matrix which is finally normed such that each row sum equals one. If the net added number of cars is *k* is positive it regards and increase of *k* cars and for *k* is negative it is of course a decrease of cars. Probabilities linked to an increase of *k* cars are placed on all the entries of the k^{th} super diagonal; if *k*<0 it would count for the k^{th} sub diagonal. The matrix is normalized such that the row sums are one. This implies of course that the transition probability matrix equals the identity matrix if there is no increase or decrease in cars.

The idea to construct the transition probability matrix this way comes from the assumption that the Markov chain is irreducible. If all transition probabilities that could not be computed from the raw data are replaced by zeros, the result might be a reducible transition probability matrix. Working with such a matrix would imply excluding realistic situations that might occur in real life. So, we simply rely on the fact that the transition behavior as expressed by an irreducible transition probability matrix at PARK200 is assumed to represent that of PARK1000.

As the Markov chain is two-dimensional, transition probability matrices are needed for each one of these dimensions of the chain. In the remainder of this section it is explained how to construct the transition probability matrix for the short-term parkers given the assumption of independency of both groups. The same procedure should be followed if one would like to find the transition probability matrix for the long-term parkers.

4.4.1 The One Step Transition Probabilities Based on the Arrival-Departure Process

The net change in the number short-term parkers results from the arrival-departure process. Therefore, this study will also consider whether it is wise to estimate the probability distribution of the net added number of short-term parkers each minute, with this arrival-departure process.

Suppose the random variable A_t is the arrival rate and D_t the departure rate of this group on time *t*. Both variables are non-negative and independent. More over A_t is Poisson distributed with distribution of the rate λ_t and D_t is binomially distributed with *n(t)*, the number of short-term parkers at time *t* and *p(t)*, the fraction of departures in this group at time *t*. The formula used to compute this fraction of departures is

$$p(t)=\frac{number\ of\ departured\ short\ term\ parkers\ at\ time\ t}{number\ of\ short\ term\ parkers\ cars\ at\ time\ t}. \quad (4.4.1a)$$

Suppose random variable Z_t represents the net number of added short-term parkers to the system at time *t* on PARK200. The net number of added short-term parkers, Z_t is then defined as A_t - D_t. The distribution of Z_t could be derived as follows:

$$P(Z_t = z) = P(A_t - D_t = z) = P(A_t = D_t + z) = \sum_{d=0}^{\infty}\sum_{a=0}^{d+z} P(A_t = a, D_t = d),$$

because of independency of the arrival-departure process, this probability equals

$$\sum_{d=0}^{\infty}\sum_{a=0}^{d+z} P(A_t = a)P(D_t = d) = \sum_{d=0}^{\infty} P(A_t = d + z)P(D_t = d).$$

As both A_t and D_t, non-negative two case can be discerned for $A_t = \text{d} + \text{z}$:
(1) If $A_t \geq D_t$, then as z=$A_t - D_t$ this implies z ≥ 0.
(2) For $A_t \leq D_t$, then z≤ 0. As z= $A_t - D_t$ and $A_t - D_t \leq 0$, it can be concluded that z≤ 0. Since both $A_t \geq 0$ or $D_t \geq 0$, *d*, the number of departures runs from –z to infinity.

$$P(Z_t = z) = \begin{cases} \sum_{d=0}^{\infty} P(A_t = z + d, D_t = d), if\ z \geq 0 \\ \sum_{d=-z}^{\infty} P(A_t = z + d, D_t = d)\ , if\ z < 0 \end{cases}, \quad (4.4.1b)$$

as A_t and D_t are independent:

$$P(Z_t = z) = \begin{cases} \sum_{d=0}^{\infty} P(A_t = z + d)\ P(D_t = d)\ , if\ z \geq 0 \\ \sum_{d=-z}^{\infty} P(A_t = z + d)\ P(D_t = d)\ , if\ z < 0 \end{cases}$$

In this way the transition probabilities $P(Z_t = k)$ can be found for having *i* cars at time *t* at the parking place and *i*+*k* cars at time *t+1*.

If the information from PARK200 should be used in the non-homogeneous Markov chain to estimate for time *t*, the transition probability matrix on PARK1000 for the short-term parkers, the parameters of the underlying arrival and departure process in $\mathcal{J}_t$ should be adjusted. Suppose that for PARK1000 the random variable L_t is the arrival distribution of the rate of the short-term parkers and M_t the departure distribution of the rate of the same group on time *t*. Then the random variable K_t = (L_t - M_t) represents the net number of short-term parkers added to the system at time *t* on PARK1000.

As it is assumed that PARK200 represents PARK1000, it's arrival rate in every unit is also representative for the whole parking place. Consider units that equal PARK200 in size. Suppose L_1, L_2, L_3, L_4 and L_5 represent the number of arrivals in five disjoint at random chosen equally sized parts of PARK1000. Then these five arrival rates, each connected to a unit, are all Poisson distributed at time *t* with the same parameter λ_t *(Ross, 2010, p313).*

Lemma 1:
Given s independent Poisson random variables, $X_i \sim Pois(\alpha_i)$ for i =1, 2, ., s, and Z=$\sum_{i=1}^{s} X_i$, then random variable Z is Poisson distributed, $Z \sim Pois(\sum_{i=1}^{s} \alpha_i)$.

The sum of *s* independent Poisson random variables is also Poisson distributed with parameter equal to the sum of the *s* associated means (See for more details appendix 3). As the arrival rates on time *t,* in the five independent equally sized units of the parking place all equal λ_t, the number of arrivals in the group of short-term parkers on time *t* at PARK1000, L_t , is Poisson distributed with parameter $\sum_{i=1}^{5} \lambda_t = 5\lambda_t$. The number of departures among the short-term parkers, M_t, is binomially distributed with *n(t),* the number of short-term parkers on PARK1000 at time *t* and *p(t),* the fraction of departures at time *t*. The fraction of departures on PARK200 at time *t* equals *p(t),* is a good estimator for the fraction of departures at time *t* on the whole parking place. The number of short-term parkers at time *t* should be estimated for the whole parking place using its probability mass function. To find the probabilities needed for that probability mass function the law of total probabilities is applied.

$$\text{P}(N_s(\text{t}) = i) = \sum_{j=0}^{1000} \text{P}(N_s(t) = i | N(t) = j)\text{P}(N(t) = j), \quad (4.4.1c)$$

where *i*=0,1, . . . , 1000 and *j*= 0, 1, . . . , 1000 respectively the number of short-term parkers and the number of parked cars at time *t*. The number of short-term parkers, N_s(t), is binomially distributed with the actual number of cars and the fraction of short-term parkers at time *t*. With the probabilities gained from statement (4.3.1b) the number of short-term parkers, *n(t),* is estimated at time *t* at the whole parking place. Using these parameters and statement (4.3.1a) the transition probability matrix can be constructed for the whole parking place of 1000 spaces.

4.4.2 The One Step Transition Probabilities Based on the Net Added Number of Cars

Suppose *N(t)* is the number of cars on time *t*, and *Z(t)* equals the net added number of cars at time *t*, then *Z(t)=N(t)-N(t-1).* To find the transition probability, p_{ij}, for having *i* cars in the parking place and in the next step *j*, implies an increase of *k=j-i* cars to the parking place. Hence the differenced series of the number of cars is used as starting point in this approach.

The probability of adding or subtracting a net number of k cars on PARK200 can be deduced from the density of the differenced series. Hereto, the smoothed version of the kernel distribution is used with all possibilities of the net added numbers per minute ranging from -200 till 200 cars. (Analogue PARK1000). If the values needed to find the probability distribution are very small and concentrated around zero, it is possible that the probability distribution turns out to be a zero vector. In that case the transition probability matrix equals the identity matrix, implying that the net added number of cars in the group of the short-term parkers is zeros in $\mathcal{J}_t$.

4.4.3 Transition Probabilities for the Higher Order Model with Triples

The transition probability matrix for the higher order model with a triple consists of the probabilities of going from one triple to another in one step. For the 201 different single states on PARK200 it would imply considering transfers for 201^3 possible triples. The dimension of a transition probability matrix that takes all possibilities into account equals 8120601 x 8120601. The whole time series uses 357 of these transfers. One could still argue that restricting the process to these known transfers, could lead to a transition probability matrix with dangling states or at least two or more classes. Hence it would be good to use the idea to allow the system to choose at random for a next state that never occurred yet. With probability θ the "regular" events do happen and with probability (1-θ), the rare events do happen. The transition probabilities could be found with a matrix *D* as follows:

$$D=\theta P+(1-\theta)S, \text{ where } \theta \in [0,1]$$

The factor θ tells the prediction algorithm with what fraction it should choose for *P*. *P* is the matrix that contains the transition probabilities of the transfers that once occurred and deduced from the probability distribution of the transfers; all other transition probabilities are set to zero. Matrix *S* contains the transition probabilities from all possible transfers at PARK200. In fact, each triple can transfer to 201 different states. The non-zero entries in matrix *S* are $(1-\theta) * 1/201$. For θ=0.85, that equals 7.4627e-04 *(Smetana, 2018, p54*).

4.4.4 The n-step transition probability matrix

As mentioned earlier in homogeneous Markov chains the *n*-step transition probability matrix P_n could be computed from the data but also estimated theoretically as the *n*th power of the one step probability matrix, P_1^n. Theoretical estimations for the *n*-steps transition probability matrix are not useful for the optimization of higher order models. Working with the theoretical approach of the *n*- step transition matrix would result in a polynomial model due to the distribution property of matrices. With the *n*-steps transition probability matrices the predicted distribution would be expressed as

$$\lambda_1 P_1\ x^{(n-1)}+\lambda_2 P_2\ x^{(n-2)} + \lambda_3 P_3\ x^{(n-3)}+ \ldots\ldots\ldots +\lambda_k P_k\ x^{(n-k)}.$$

If the Kolmogorov equation holds this linear relation would be:

$$\lambda_1 \, P_1 \, x^{(n-1)} + \lambda_2 \, P_1^2 x^{(n-2)} + \lambda_3 \, P_1^3 x^{(n-3)} + \ldots\ldots\ldots + \lambda_k \, P_1^k x^{(n-k)},$$

implying that every term depends on the one-step transition probability matrix. Optimization of the latter will always be with respect to the one-step transitions such that the direct influence of the *n*-step transitions is might be neglected.

Homogeneity is assumed in an interval, but to find the right interval where homogeneity really counts is not always clear. Therefore, the *n*-step transition probability matrix will be computed with the differenced series of the net added number of cars after *n*-steps. This implies that the *n*-step transition probability matrix can be found by applying the reasoning in section 4.4.2 on lag *n* such that the differenced series equals *Z(t)=N(t)-N(t-n).*

5. Application of Markov chain prediction models

Markov chain models do lack accuracy in predicting temporal data. Therefore, this section provides an algorithm that take the existence of two dimension in the group of parkers and the temporal evolution of the number of parkers into account. It combines Markov chain prediction models together with mathematical techniques in order to solve the mathematical problem as described in section 3.4. In making predictions, the algorithm assumes that the part of the Markov chain in $\mathcal{J}_t$ is homogeneous and that therefore the transitions for that part of the chain can be used to estimate the transition probability matrix at time *t.* After introducing the algorithm in the first part of this section, the second part explains how and when Markov chain prediction model will be used in the algorithm. In doing this it also shows what mathematical techniques will be used to improve the results of Markov chain prediction models. The discussion ends with explaining how the prediction results will be evaluated.

5.1 The algorithm

The algorithm aims to predict the number of cars on PARK1000, the complete parking place, at time *n* using the number of both groups at time (*n-s)* with the “known” data from PARK200, a set of 200 parking spaces equipped with sensors on the parking space. Hence, it is tried first to train the data from PARK200 and to find a model that works “good” on PARK200. This model is then tested with the data from PARK1000.

To achieve that goal a basic model is applied iteratively. The main idea of this basic model is to predict the number of cars on PARK1000 for time *n* starting at current time *n-1*:

1) Compute the probability distribution vector for both dimensions or groups separately using a Markov chain prediction model.
2) Compute the distribution of the number of cars using the modified convolution.
3) Calculate the expected number of cars and the ranges for the prediction interval.

Predictions starting from time *n-1* till time *n+s, s*>1, are done iteratively with the basic model.

Schematically the algorithm consisting of three blocks reads:

I. Given the initials at time *n*-1:

$\mathcal{J}_{n-1}$ = time interval that isolates a part of the Markov chain that will be used to estimate the transition probability matrix at the actual time

k = the number of historical time points that are involved in the model, k=1, 2, ……, length ($\mathcal{J}_{n-1}$)

II. Determine for each group (dimension) at time *n*-1:

- the initial distributions at each historical time point.
- the *n*-steps transition probability matrix for the homogeneous part of the chain in $\mathcal{J}_{n-1}$
- the optimal Markov chain prediction model

III. Compute for time *n*

- the predicted distribution for each group separately
- the probability distribution for the number of cars with the convolution concept
- the for random noise corrected predicted distribution by using a convex combination

Technically the whole algorithm reads:
Given the time dependent stochastic processes of the short- and long-term parkers at time *t*, on day *d*, *X(t)* and *Y(t)*, where *t*=0,1,2,…..(*n-1*), then the probability distribution of the sum process for the next minute $u^{(n)}$equals a for noise corrected modified convolution of the predicted distributions of each one of the basic processes.

The modified distribution $u^{(n)}_{modified}$, follows from the convolution, indicated in this report as *,

$$u^{(n)} = \sum_{j=1}^{k}(\lambda_j \quad x^{(n-j)} P_{j,\mathcal{J}_{(n-j)}} \ * \ \sum_{j=1}^{k}\theta_j \quad y^{(n-j)}Q_{j,\mathcal{J}_{(n-j)}} \, , \tag{5.1.1}$$

and is corrected for random noise by using a convex combination of the predicted and a next distribution function $v^{(n)}$:

$$w^{(n)}=\beta u^{(n)}_{modified} + (1-\beta)v^{(n)}, \tag{5.1.2}$$

where $v^{(n)}$ represents the probability distribution on time *n* for the average rate on PARK1000 aggregated by time and day, $\beta \in [0.1]$. $w^{(n)}, u^{(n)}_{modified}, v^{(n)} \ \in \mathbb{R}^{1001}$. In the remainder of this report that next distribution, $v^{(n)}$ will be referred to as a corrector.

For the first order model statement (5.1.1) is simplified to: $u^{(n)} = \ x^{(n)} * \ y^{(n)}$

, where $x^{(n)} = x^{(n-1)}P_{1,\mathcal{J}_{(n-1)}}$ and $y^{(n-1)} = y^{(n-1)}Q_{1,\mathcal{J}_{(n-1)}}$.

5.2 Discussion of the Prediction Algorithm

The first part of this section discusses the three blocks of the algorithm in the light of the Markov chain prediction model that is used to do the actual predictions for each dimension. Higher order models with combo states are very laborious, due to the large state space, and will be approximated by the higher order model as proposed by Chin (section 4.3). The second part of this section considers two ways the algorithm is applied in this research.

5.2.1 *The Three Blocks of the algorithm*

Generally, predictions will be done from scan moment to scan moment. Given the current date and time, a time interval is determined such that a part of the non-homogenous Markov chain is "declared" homogeneous. As mentioned in section 3.2 three types of intervals are explored in this research: 1) eight parts or intervals for each day, 2) disjoint consecutive parts of *h* minutes each day and 3) a moving interval of *h* minutes each day. In the first order model one historical point is considered and in the higher order models more than one. As it is assumed that time interval $\mathcal{J}_t$ is needed to find $P_{\mathcal{J}_t}$, the number of historical timepoint, *k,* should be chosen such that the last historical point still lies in that time interval. The heuristic that is used to determine the number of historical points is: There are at least *k* points needed to express a "polynomial" relation of degree *(k-1).* The degree of the polynomial relation can be seen from the form of the plot of the number of parkers in each group by time.

In block two of the algorithm, the initial distribution(s), the transition probabilities and the optimal model are determined. The set of initial distribution for a model that takes *k* historical points into account, counts for each group of parkers *k* initial probability distribution vectors each one linked to one historical point. The initial distribution on PARK1000 is estimated by using the data associated with time *t.* (See section 5.2.2 for more explanation on the choice of the initial distribution).

The transition probabilities are found using the net added number of cars each minute. The parameters needed to find the distribution of this series should be estimated, except at the time point *(t-1)* that the prediction algorithm starts a run. As section 3.3 concluded the transition probabilities needed for time *t*, are estimated with data coming from, $\mathcal{J}_t$, a time interval containing time *t.* This includes that a user of the algorithm constantly must take the actual time and actual day into account while doing predictions. Estimation has to be done separately on each dimension of the Markov chain. The discussion will be linked to that of the short-term parkers. But it is fairly similar for long-term parkers.

For finding the transition probability matrix the average of the net added number of short-term parkers is aggregated by time and day. For the matrix based on the arrival departure behavior of short-term parkers, an estimation is done for the arrival rate, and the parameters of the departure distribution for the short-term parkers. The arrival rate on time *t*, λ_t, is estimated by $\hat{\lambda}_{\mathcal{J}_t}$, the mean arrival rate for the time interval $\mathcal{J}_t$, such that the distribution of the arrival rate in that time interval (by expansion similar time intervals) by the assumption of homogeneity is Poisson distributed with parameter $\hat{\lambda}_{\mathcal{J}_t}$. The distribution for the departure rate is binomially distributed on time interval, $\mathcal{J}_t$. The fraction of departure within the group of the short-term parkers as given in (4.4.1a) is estimated by, $\hat{p}_{\mathcal{J}_t}$, the mean fraction of departures per minute on that time interval. Hereto the number of departed short-term parkers and the number of short-term parkers are each aggregated by day and by time, such that the fraction of departures per minute can be estimated at time *t.*

The parameter for the first order model that uses one historical time point is trivial. The parameters for the higher order model as proposed by Chins are optimized as explained in section 4.3. It is believed that the stationary distribution represents the distribution of the states on the long run, such that it could be seen as a good estimator for the state distribution of the Markov chain. The algorithm of the power iteration is used to find the stationary distribution. As this algorithm uses a starting point at random to approximate the eigenvector with the largest eigenvalue one, it does not always find an eigenvector. In that case all the parameters are equal.

In the third block of the algorithm, distributions for the next minute are predicted separately for both dimensions of the Markov chain. For the first order model the basic prediction model is chosen. For the higher order model it is chosen to explore the higher order model as proposed by Chin et. al. In the higher order model with the triple of states the probability distribution for transferring from triple one at time *t-1*, to triple 2 at time *t*, is in fact a joint distribution from triple one and triple 2. The predicted distribution is the marginal distribution of triple 2. Due to the large space of transition That model uses the optimal linear combination of the distribution of the historical points in the combo to predict the next combo.

If the algorithm uses the first order Markov chain model to do the actual predictions, it will be referred to as Markov2D and HomcChin2D if the algorithm uses higher order Markov chain prediction models to do the predictions each minute. The probability distribution for the total number of cars (both the short-and the long-term parkers together) is computed with the convolution concept. The convolution gives the probability for every possible sum of the number of short- and long-term parkers, but the maximum number of cars that can be parked is limited. It is not clear whether the cars that could not be parked when the parking place is full are really lost "customers". That is why the probabilities on these events are included in the probability of the parking place being full. In this way the convolution is somewhat modified such that it can be used as probability distribution for the number of occupied parking spaces.

The algorithm prescribes correcting the predicted distribution for noise every minute. Corrections are always done on PARK1000. The number of cars at PARK1000 are binomially distributed with N and p. The average fraction, a, of cars on PARK200 aggregated by day and time is a good estimator for the occupation rate on PARK1000. So, the probability distribution on time t, can be estimated using a binomial distribution with parameters N and a. The convex combination that gives an even weight to the predicted distribution and the distribution deduced from the average fraction of the cars on time t is used to rectify prediction "flaws". In this study the default value for β in statement (5.1.2) is 0.5. For $\beta = 1$, one would simply rely only on the predicted values as result, and $\beta = 0$, would mean just working with the average what is probably a good estimator but not sensitive to future changes in parking behavior.

Markov chains predict the expected value of the number of cars at the parking place. But as the data is coming from a time series the realization depends mainly on the time, displaying sudden changes in reality. For such changes it will be considered if it is better to use another estimator to determine corrector the distribution vector deduced from another estimator for the number of cars. Under the assumption that the third quartile, K_3, is a good estimator for the number of cars, the distribution of the number of cars is binomially distributed with the number of parking places N, and fraction $\frac{K_3,}{N}$. The same reasoning can be followed for choosing to correct with another distribution, based on the 95th percentile.

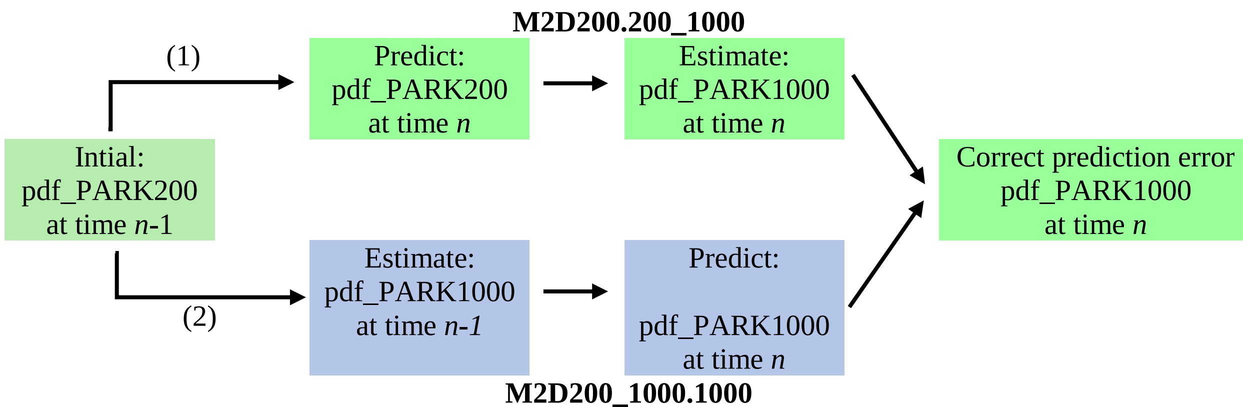
The input for the next iteration is the resulting for noise predicted distribution on time *t*.

5.2.2 Two approaches for applying the algorithm

In this research the number of cars at the parking place within each dimension can be predicted in two ways with the algorithm such that there are two distinct ways of applying this model: 1) M2D200**.200_**1000 and 2) M2D200**_1000.**1000. The big difference is the way the set of initial distributions is chosen. This is explained for the group of short-term parkers. (Analogue for long-term parkers)

In the first approach (M2D200**.200_**1000) the number of short- and long-term parkers on PARK200 (the first 200 spaces of PARK1000) is used to find the initial distribution for each group at PARK200. The initial distribution vector for short-term parkers contains the probabilities of the process for being in that state at time *n*. Since the state is known, the probability of being in that state equals one. Hence, the initial distribution is a zero-row vector with a one at the entry representing the current state. In case the number of short-term parkers is known and equals *j* (*j*=0, 1, . . . , 200), entry number *(j+1)* of this vector is one and the other entries zero, $\bar{x}_j^{(n)} = ((0, 0, \ldots, 0, 1, 0, \ldots, 0)^{\mathrm{T}}$ (*Liu Tie*, *2010*, *p163*).

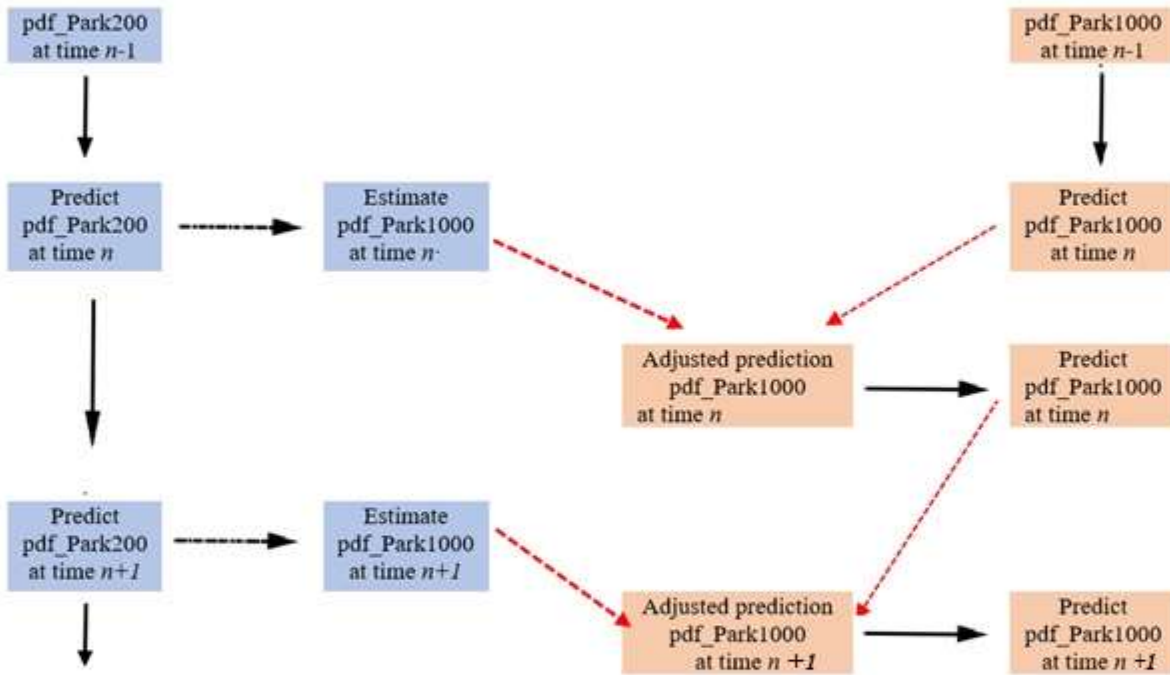
This initial distribution of PARK200 consists of 201 entries and is used to predict the distribution of the number of short-term parkers at the next minute on PARK200 using a Markov chain prediction model. Each minute a prediction is made for each group separately, the convolution is computed and modified for PARK200. The expected predicted number of cars is computed and used to find the expected predicted fraction of cars on PARK200. This fraction is a good estimator for the fraction of cars on PARK1000 and therefore used to find the expected distribution of the number of cars on PARK1000. So, in this approach predictions are done for PARK200 and the predicted fraction or the expected rate on PARK200 is used to estimate the distribution of the occupation rate at PARK1000.



The second approach (M2D200**_1000.**1000) estimates first the distribution of the occupation rate on PARK1000 at time *t* with data from PARK200, before doing any prediction. The average fraction of short-term parkers aggregated by time and day, $p(t)$, is a good estimator for the fraction of the similar group at PARK1000. The number of short-term parkers on PARK1000 is binomially distributed with *n(t)*, the number of parkers and *p(t)*, the fraction of short-term parkers within the group of parked cars at time *t*. (*n(t)*=0, 1, 2, . . . , 1000). These parameters are used to estimate the initial distribution for short-term parkers at PARK1000 such that predictions can be made for this group on PARK1000. (Analogue the long-term parkers.) The number of cars at PARK1000 is computed with the convolution of the distributions of both groups. In this approach predictions for PARK1000 are done with the estimated initial distributions on PARK1000 for both groups separately. The number of cars is found by applying the convolution concept on the predicted distributions.

For both approaches, the for noise corrected predicted distribution is used to find the initial distribution needed for the next iteration.

Moreover, it was considered also whether these two approaches could be combined to one model. In the diagram below, it can be seen that predictions are done with the first approach, and predictions are done for PARK1000 with data from PARK1000 (data retrieved from the scan vehicles). A weighted distribution for both prediction results is used as input for the next prediction starting from PARK1000. While the predictions on PARK200 are proceeded without adjusting its result.



5.3 Model evaluation

The main standard for evaluation of the models in this research is the prediction accuracy in a prediction interval at significance level alpha (in this study 5%) (*Liu Tie*, *2010*, *p164*). The prediction accuracy equals the percentage of real values that lie in the prediction interval. The prediction accuracy will be measured for predictions in the first *s* minutes and in the s^{th}minute. A good model should also have a reasonable mean range for the prediction interval (See section 2.1) and a reasonable running time. It's a posteriori error should display a random effect, as randomness is a main component of a time series. The aim is to filter out a model that excels also in the percentage of good predictions both in a time period of *s* minutes and in the s^{th} minute.

The 95% prediction interval concerns an interval for the for noise-corrected prediction rate on PARK1000 as resulted from (5.1.2). The ranges of the prediction interval can be found empirically or theoretically. Once a predicted distribution is found "that is considered to be good", it is known which probability belong to which state. By filtering for the states where the cumulated sum of probabilities is respectively less than $\frac{\alpha}{2}$, and (1-$\frac{\alpha}{2}$) the lower and the upper range of the prediction interval can be found.

Theoretically, the ranges for the prediction interval of the distribution of the occupation rate are:

$$\hat{p} \pm t_{\frac{1}{2}\alpha,(n-k+1)}\sqrt{\frac{\hat{p}(t)*(1-\hat{p}(t))}{m^2}+\frac{\hat{p}(t)*(1-\hat{p}(t))}{n}},$$

Predictions are done from a starting point t=t_0 an end point t=T. This is done several times, each time starting at another start point. For simplicity a time from start to end point is referred to in the remainder of this report as a run. For each prediction it is checked whether the real measured test data lies within the 95% prediction interval resulting from the predictions gained from the model built with the training set. Hereto the prediction accuracy in a period of *n* minutes in the j^{th} run is computed as:

Prediction accuracy in *n* minutes= $\frac{100\%}{T-t_0} \sum_{t=t_0}^{T} a_{t,j}$, (n=T-t_0)

and $a_{t,j} = \begin{cases} 1, & if\ occupation\ rate\ lies\ in\ prediction\ interval\ on\ time\ t\ in\ run\ j \\ 0, & otherwise \end{cases}$

After completion of *j* runs (*j=1, 2, . . . ,J)*, the prediction accuracy can be computed in the n^{th} minute.

Prediction accuracy in the n^{th} minute= $\frac{100\%}{J} \sum_{j=1}^{J} a_{n,j}$,

and $a_{n,j} = \begin{cases} 1, & if\ occupation\ rate\ lies\ in\ prediction\ interval\ on\ time\ n, in\ run\ j \\ 0, & otherwise \end{cases}$

The null hypothesis that the prediction accuracies equals 90% implies that it is assumed that 90% of the test data lies between the ranges of the 95% prediction interval. If the null hypothesis is not rejected at significance of 5%, it is thought that the data does not provide enough evidence to conclude that the occupation rate of the test file does not predicts the occupation the rate of the training file.

As mentioned in section 2 the predictions of the training model will be adjusted after doing and correcting the prediction for noise. The model is able to predict the scan data if the rate retrieved from the scan data lies in the adjusted prediction interval. Hereto different relations between the scan rate, S_t, and the sensor rate, R_t, and the difference of the sensor rate and the scan rate, ε_t, on time *t* are used:

- $S_t = aR_t + b$. The most general way to estimate data is by using the regression line, or line in the "middle" of the scatter plot, such that we have least sum of squared errors. This is only valid if the scatterplot reveals a linear relation between variables.
- $\varepsilon_t = \sum_{i=0}^{n} a_i R_t^i$. A polynomial function to approximate the periodic behavior. (See figure 2.1a)
- $S_t = R_t$-ε_t. Basically, the a-priori error, ε_t, equals the occupation rate from PARK200 (R_t) minus the occupation rate from PARK1000 (S_t). This error is corrected afterwards. But as the actual values are unknown, estimators will be used to approximate the error. $S_t \approx \hat{R}_t$-$\hat{\varepsilon}_t$, where $\hat{R}_t$ equals the for noise corrected predicted rate and $\hat{\varepsilon}_t$, a local measure for the error. The most preferable one is the median, since there seem to be outliers. Nevertheless, the mean and the mean mode of the a-priori error linked to the intuitive partitioning of eight periods will be also considered.

The randomness of the predicted values will be examined using the a posteriori error or Predicted rate minus Scan rate. If the a posteriori error of a model with a high prediction accuracy is normally distributed the model can be called good (*Coghlan A., 2010).*

6 The Research Results

This section contains interesting results that could help to determine which model performs better and to what extend it could be relied on. The evaluation of the prediction results will be mainly focused on the prediction accuracy. First the results for the different varieties of Markov2D algorithms will be discussed. Out of these Markov2D algorithms the variant with the best performance will be chosen. The second topic that will be evaluated is the application of the HomcChin2D algorithm using the best variant of the Markov2D algorithms as base. It is important to remember that the higher order Markov chain models are introduced to examine whether the performance of the best variant of the first order model could be improved. Hence higher order models will not be explored with all varieties of applying a Markov chain prediction model in this study.

6.1 Evaluation of the Markov2D algorithms

The algorithm is first explored with the two approaches mentioned in section 5.2.2. For both groups the sub models are evaluated with transition probabilities based on the arrival-departure behavior (*AD*) and the fit distribution of the net added number of cars (*Fit*). Moreover, the models are all explored with the three ways of choosing a time interval in which homogeneity is assumed (*Eight*, *Fix* and *Moving* periods). These different aspects imply the evaluation of 12 sub models. In this research the models are explored for fixed periods of 30, 15, 10, 5, 3 and 1 minute(s). The best results were retrieved by using periods of 15 minutes. It seemed like the smaller the time span the more random noise the prediction gets. Granularity does not automatically improve prediction accuracy of the model. Transition probabilities needed in these sub models are found with 1) the whole history of similar time intervals and 2) a good estimator for the transitions in the actual time interval.

To choose the best trained model on PARK200, the prediction results of 10 runs are compared. One run starts to predict from scan moment to scan moment. The intermediate results were also evaluated to compare the models. One way to find the start moment of a run is by selecting at random 10 scan moments. A next way is to select the first scan moment of a group of 10 consecutive scan moments at random. The latter is used for comparing all 12 sub models. The results of training the Markov2D algorithms are summarized in the table here below. Table 6.1a contains the prediction accuracies, the mean range of the prediction interval and the CPU time.

Table 6.1a: Performance of the M2D200_200.1000 models using whole history or a good estimator on PARK200

Model	All previous transitions history			Good estimator		
	Prediction accuracy (%)	Mean range	CPU time (s)	Prediction accuracy (%)	Mean range	CPU time (s)
	10 consecutive runs					
M2D200_**200.**1000AchtAD	52.91	0.037	3746	52.95	0.037	2743
M2D200_**200.**1000FixAD	53.08	0.037	2805	53.19	0.037	2585
M2D200_**200.**1000MovAD	53.13	0.037	4639	53.13	0037	2604
M2D200_**200.**1000AchtFit	53.23	0.038	173	52.74	0.037	155
M2D200_**200.**1000FixFit	53.06	0.038	144	52.91	0.037	137
M2D200_**200.**1000MovFit	53.10	0.038	143	**53.06**	**0.037**	**136**
	10 runs at random					
M2D200_**200.**1000AchtAD	58.44	0.037	6281	58.16	0.038	2939
M2D200_**200.**1000FixAD	57.67	0.037	6076	58.52	0.037	3768

M2D200_**200.**1000MovAD	59.75	0.037	4687	55.71	0.037	2558
M2D200_200.1000AchtFit	57.60	0.038	1987	58.77	0.037	249
M2D200_200.1000FixFit	60.33	0.038	655	59.57	0.037	213
M2D200_200.1000MovFit	61.53	0.038	293	**62.11**	**0.037**	**190**

On this point of the study it is checked whether relaxing on the assumption of needing the whole history to determine the transition probabilities is possible. For both cases the prediction accuracies a noted in the Table 6.1a were compared. A t-test of paired observation that assumes equal variance was run to compare the prediction accuracy resulting from the models in above mentioned. With p-value 0.2921 this test convinces us that it cannot be said that the means are not equal (A t-test for the results of random sampling gave a similar result, p-value=0.7062). From this point on the heuristic will be used that the mean of respectively the number of cars, arrivals and departure are good estimators to determine the associated transition probabilities (See introduction of chapter 4). For both ways of sampling it is clear that the sub model in this group, M2D200_200.1000, that uses a fit distribution and assumes existence of a moving homogeneous Markov chain performs the best in prediction accuracy and CPU time.

The performance of the M2D200.1000_1000 sub models are evaluated. The results give the same picture: Better performances for models using the "slowly forward moving intervals" and the fit distribution.

Table 6.1b: Performance of the M2D200.1000_1000 models on PARK200

Model	Prediction accuracy (%)	Mean range	CPU time (s)
	10 consecutive runs		
M2D200.1000_1000AchtAD	26.94	0.079	31073
M2D200.1000_1000FixAD	67.43	0.077	44370
M2D200.1000_1000MovAD	67.80	0.078	34476
M2D200.1000_1000AchtFit	51.11	0.078	815
M2D200.1000_1000FixFit	49.76	0.078	757
M2D200.1000_1000MovFit	53.17	0.078	742
	10 runs at random		
M2D200.1000_1000AchtAD	84.55	0.061	67201
M2D200.1000_1000FixAD	84.67	0.060	47162
M2D200.1000_1000MovAD	84.45	0.061	43154
M2D200.1000_1000AchtFit	84.26	0.060	1337
M2D200.1000_1000FixFit	84.19	0.060	905
M2D200.1000_1000MovFit	86.56	0.060	753

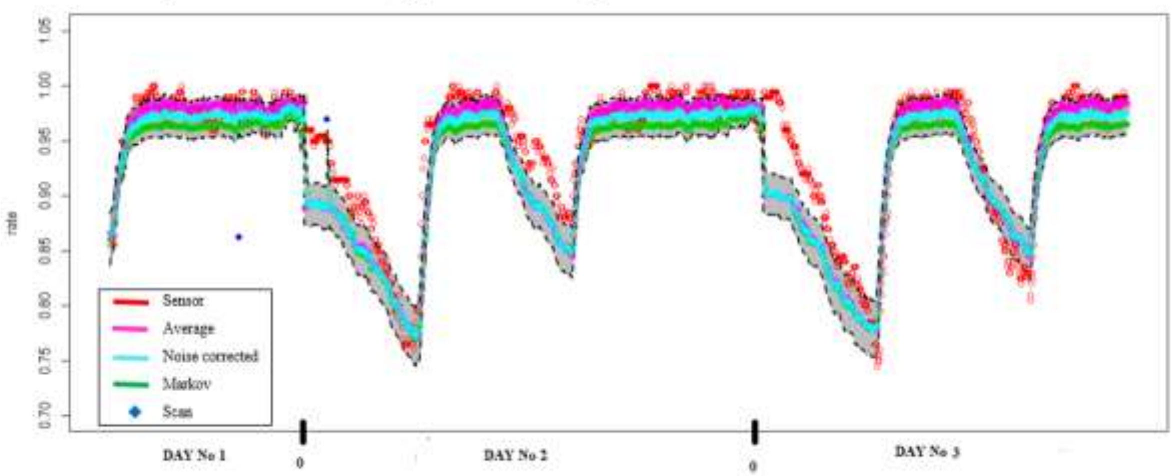
The models based on the second approach do have a larger CPU time. One reason is that each prediction done is of order N^2 as it includes working with a transition probability matrix with N states. The larger the dimension of the matrix, the more time needed to finish a single prediction. Add to this all computations needed for finding the transition probabilities for the Poisson and the binomial distribution. Including the arrival-departure behavior into the model does not result necessarily into a better model. In this approach the percentages are higher, but the mean range exceeds the ideal range of 0.043 as deduced from the a-priori error in this report.

A combination of both approaches in one model as mentioned in section 5.2.2 gives for the same groups of consecutive runs a prediction accuracy of 46% and mean range 0.061 (CPU=1530sec). Combining the two approaches does not necessarily lead to a better model. Hence M2D200_200.1000MovFit is chosen as the best of all the 12 Markov2D sub models discussed in this study.

6.1.1 Performances of the Best Model, M2D200_200.1000MovFit

In this section the focus will be on important results from the effort to improve the performance of the best model, M2D200_200.1000MovFit on PARK200. The predictions tend to “behave” like the sensor rate. To give the reader an idea the predictions for the first 1600 minutes are plotted in figure 6.1.1a. The green graph are the prediction results gained from the convolution. The red graph represents the real occupation rate that the algorithm tries to predict. As the figure shows, the predicted rate is most of the times less than the sensor rate. See for more details figure 6.1.1a.

Figure 6.1.1a: Different type of rates in the prediction interval of the for noise corrected rate



M2D200.200-1000MovFitConvMod 53.06%, 0.037mean range, 136sec, K=10

From the graph it is clear that the model fails to predict the first minute(s) of a day. The real rates are most of the time under estimated by the predicted rate; the predicted rate as result from the convoluted prediction with the Markov chain model is most of the time under the real value of the sensor rate. Generally, the average rate aggregated by time and day is closer to the sensor rate then the predictions. The blue graph is the for random noise corrected prediction, resulting from the convex combination (5.1.2). The noise corrected rates do improve the predictions, but they cannot do better than the a fore mentioned average rate.

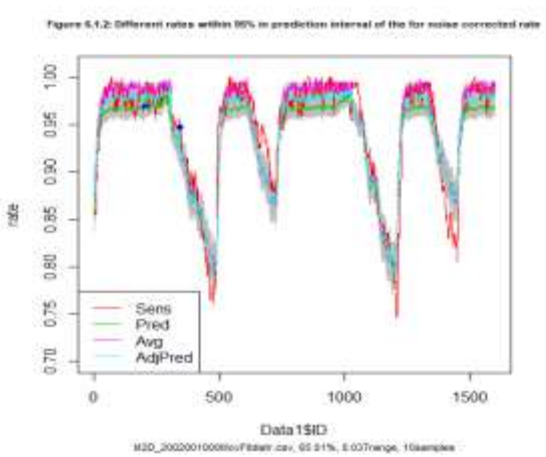
For the best model M2D200_200.1000MovFit the prediction accuracy of 100 consecutive samples equals 58.10% and the mean range 0.040. The prediction accuracy is also checked in the n^{th} minute and for the first *n* minutes of the prediction. High percentage can be expected in the first minute of the sample. In the 5^{th} minute 69.87% of the points were correctly predicted and in the 15^{th} minute 61.05%. The prediction accuracy for the first 5 minutes equals 67.53% and for the first 15 minutes it equals 60.20%.

6.1.2 Improving the Performance of M2D200_200.1000MovFit

It will be examined whether the performance of best variant of the sub model thus far, can be further trained to improve the prediction accuracy by relaxing on some assumptions. It regards assumptions of existence of two main groups at the parking place, homogeneity in an interval containing time *t*, irreducibility of the transition probability matrix and the assumption that the mean sensor rate of the known data aggregated by day and time is a good estimator for the sensor rate. Results are generated and compared for the same periods of hundred consecutive runs. In this section, the research results from training the model further on PARK200 can be found.

It was considered whether the best model could be simplified to a *one-dimensional (1D) model* such that it is not needed to take the behavior of short- and long-term parkers separately into account. The prediction accuracy of the 1D-model in the similar time periods equals 51.56% and the mean range equals 0.028. The period with prediction accuracy 90% is on average 11 minutes and the prediction accuracy in a period of 30 minutes equals on average 54.40%. The prediction accuracy of the 2D-model counts 58.09%. The mean time with prediction accuracy of 90% equals 14.9minutes and the prediction accuracy in a period of 30 minutes equals on average 58.94%. A t-test for paired runs shows that at significance level 5% the data does not let one conclude that the prediction accuracy of the 2D-model is not higher than the prediction accuracy of the 1D-model. So, it cannot be expected that working with one dimension will lead to a better model.

Transition probabilities at time *t* are linked with the associated “moving time interval”. The last minute of the day (*t*=719), however is a special moment (section 1.3.5). For predictions at the end of the day it was considered whether *restricting the “moving time interval” to the first minute* of the day would improve the model. A t-test reveals that the improvement in correct predicted values is neglectable (p-value=0.217). *Relaxing on the assumption of irreducible* transition probability matrices would imply that only the transition probabilities for states ranging from the minimum to the maximum number of cars included in the transition probability matrix; all other probabilities are set to zero. A t-test of paired means convince us that it cannot be concluded that the prediction accuracy of the model with reduced transition probabilities is not less than that of the best sub model (p-value= 0.0007). The prediction accuracy for the model that excludes not visited states equals 57.51% and mean range 0.035.



To understand the importance of using a corrector as suggested in 5.1.2, one should realize that using the Markov chain prediction models alone (β=1) resulted the median of the prediction accuracies of 35% on PARK200. As it is clear from the graph and the role of convex combinations, *the way the predicted* values are corrected from noise can perform as best as the average itself. The average does not approximate the real values accurately in every interval. Although it “brings” the predicted value resulting from the convolution closer to the real values, the model still has a lot of predictions less than the real values.

So instead of trying to find a better weight for the convex combination between the average and the predicted value, it is considered to choose another distribution vector as “corrector”. The next “correctors” that are used are the distribution vector of the of the median, the 75th and 95th percentile. The best results on the same interval are found at with a distribution based on the third quantile as estimator of the number of cars. It resulted in a model that could predict the rate at the first minute of the day. The prediction accuracy equals 64.78% and a mean range of 0.037. A combination of relaxation of the moving interval at the end of the day and correcting prediction flaws with a distribution based on the third quantile leads to a better model, such that prediction accuracy equals 65.01% and mean range of 0.037. Higher percentiles do not automatically lead to higher percentages. Noteworthy is that the higher the percentile that is used to find the distribution of the corrector, the less the model is able to predict lower sensor rates.

This brought up the idea to use two “correctors” to reduce prediction flaws: the distribution based on a low and a high percentile as estimator for the expected value. Two combinations were explored: first and the third quartile and the 5th and the 95th percentile. On PARK200 the prediction accuracy equals 77.52% and the mean range 0.056 when using the distribution of the first and the third quartile to correct for prediction flaws. Using the distribution of the 5^{th} and the 95^{th} percentile to correct for prediction flaws gives a prediction accuracy of 95.79% and the mean range 0.092 on PARK200.

Including the correction with the distribution of the 25^{th} and 75^{th} percentile in the algorithm gives a prediction accuracy of 92% and the mean range 0.058 on PARK1000. The CPU time equals 7407seconds. Correcting for prediction flaws with the distribution of the 5^{th} and the 95^{th} percentile resulted in a prediction accuracy on PARK1000 is 100% and the mean range equals 0.091. These adjustments are not seen as improvements. The a posteriori error are not normally distributed According to the Shapiro-Wilk normality test it cannot be said that the a posteriori error resulting from the model including correction with quartiles, comes from a normal distribution (p-value =0.0000). The prediction accuracy is improved if the corrections of the 5^{th} and the 95^{th} percentile are included in the algorithm, but the mean range is far from ideal.

Hence the Markov2D algorithm that includes the fit distribution, the moving interval, the focus on the last minute of the day and the correction with the 75^{th} percentile, will be considered as the best model thus far. It is able to predict on average a period of maximum 50.5minutes 90% of the data accurately. In the first 30 minutes it prediction accuracy equals 71.62% on PARK200. In the remainder of the study it will be referred to as M2D200_200.1000MovFitNightQuart3.

6.1.3 Adjustment to Model to Predict the Scan Data

The model thus far, coded by: “M2D200_200. 1000MovFitNightQuart3” is able to predict 73% of the occupation rates on PARK1000. At this point is was checked if correcting the predicted value with the a priori error would lead to better predictions. Section 5.3 mentioned several ways to do that. In this research the linear relation and the polynomial function were not explored intensively in this research. As far as it could be implied the linear relation ($S_t = 0.9774R_t + 0.0201$) did not lead to reasonable results. The best fit polynomial relation is unstable and did not improve the predictions accuracy.

$$(\varepsilon_t = 10^{17} * \left(-0.0552R_t^8 + 0.3837R_t^7 - 1.1653R_t^6 + 2.0194R_t^5 - 2.1843R_t^4 + 1.5102R_t^3 - 0.6518R_t^2 + 0.1606R_t \ - 0.0173\right)$$

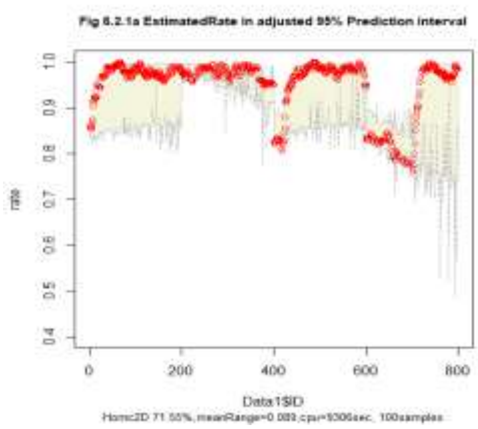
The third option, S_t=R_t-ε_t, for adjusting the outcome of the model such that the scan rate could be approximated was explored. It captures the idea of the a-priori error. In each one of the eight parts of the day the error is estimated by the mode (in case of more, the mean mode), the mean and the median based on historic data. Within a 95% prediction interval the prediction accuracies are respectively 54%, 56% and 58% with mean range 0.026, 0.027 and 0.027. The a-posterior error gives also the same picture as the a-priori error and is right-skewed. Adjusting the prediction afterward with the a-priori error such that it approximates the scan data does not lead to improvements.

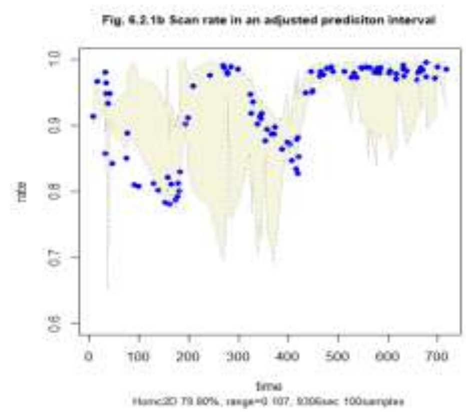
6.2 The Higher Order Model HomcChin2D

In this section it will be considered whether the best model, M2D200_200. 1000MovFitNightQuart3, could perform better by replacing the basal Markov chain prediction model with the higher order model of Chin et. al. The latter is explored with 5 historical time points for the short-term parkers and 3 historical time points for the long-term parkers. As it regards improving the model this part of the research will focus only on the main criteria of improvement: the prediction accuracy and mean range for both PARK200 (sensor rate) and PARK1000 (scan rate).

6.2.1 Results from Applying HomcChin2D to the Best Model

To see whether HomcChin2D could lead to better performances the focus will be on predicting a scan moment starting 200 minutes earlier. The 100 consecutive samples starting to predict the scan value in line 134 from the Scan file, resulted after 9862seconds in a prediction accuracy of the sensor rate of 71.55% and a mean range of 0.089. This model is able to predict 79.80% of the scan data and the mean range equals 0.107.





Although a higher mean range is not desired, it is needed to capture the large transfers of the occupation rate each minute. Interesting is also the fact that the model has difficulties to predict the scan data in the first part of the day and the lowest values for the scan rate. The result is now a sub model which will be referred to as: HomcChin2D200_200. 1000MovFitNightQuart3.

The performances with regard to time was also examined. The maximal period of minutes that the model had a prediction accuracy of 90% equals on average 63.4minutes. For the HomcChin2D the performance of prediction accuracy is examined after n minutes and in the n^{th} minute of the run. In table 6.2.1a the result for the first 10 minutes can be found.

Table 6.2.1: Mean Prediction accuracy after n minutes and in the n^{th} minute of the run

n	Prediction accuracy		n	Prediction accuracy		n	Prediction accuracy	
	After n minutes	In the n^{th} minute		After n minutes	In the n^{th} minute		After n minutes	In the n^{th} minute
1	1.00	0.33	6	0.90	1.00	11	0.86	1.00
2	0.96	0.67	7	0.91	1.00	12	0.86	0.67
3	0.93	0.67	8	0.87	0.33	13	0.86	0.67
4	0.92	1.00	9	0.86	0.33	14	0.83	0.67
5	0.91	1.00	10	0.85	1.00	15	0.85	1.00

The prediction accuracy itself is a time series. For predictions after *n* minutes it can be said that the higher the prediction time, the lower the prediction accuracy until it seems to fluctuate around 77.80%. Although the prediction in the n^{th} minute could be reliable, that is not the case for every minute. Therefore, it will be considered in the next section whether the prediction accuracy of this model could be improved.

6.2.2 Improvements for Model HomcChin2D200_200. 1000MovFitNightQuart3

This section presents the results of the efforts made to improve the performance the prediction accuracy for PARK1000. The next adjustments regard improvements for both the CPU time and the prediction accuracy. First the result of relaxation of the assumption of the existence two independent groups at the parking place and reducing the number of historical points in the higher order model are presented. Then the effect on deterministically choosing the parameters for the prediction model is explored for three set of parameters. By relaxing on these three assumptions it was considered whether at least the same prediction accuracy could be reached while the CPU time is reduced. Finally, it was examined whether the idea of two correctors could be used to correct prediction flaws.

The HomcChin2D algorithm was explored for a *one-dimensional Markov chain*, and resulted in a lower prediction accuracy of 77%, mean range 0.107 although the CPU time was reduced up till 5747sec.. Reducing the number of lags was explored for two cases. The sub model with 4 historical points for the short-term parkers and 3 for the long-term parkers resulted in a prediction accuracy of 78.79%, mean range 0.107, CPU time=7746sec.. For 3 and 2 historical points for respectively the short- and the long-term parkers the prediction accuracy was 74.49%, mean range 0.107, CPU time 5744sec and a posteriori p-value 0.07865. Although the CPU time decreased because of these adjustments, the prediction accuracies are lower than that of the best sub model up till now. Hence, these adjustments are not classified as improvements.

As statement (4.3.3) suggested the *optimal set of parameters* in the HomcChin2D should be chosen such that the error with regard to the stationary distribution is minimal. Relaxation on this assumption was done and the parameters of each group was chosen such that the contribution of every historical point was given a weight that represents the contribution to the distribution to the prediction. Some of the results of deterministically choosing parameters can be found in the table below.

Table 6.2.2a: Prediction accuracy and mean range for scan data by set of model parameters

Model Parameters		prediction accuracy	mean range	CPU time seconds
Short-term parkers	Long-term parkers			
Optimal set	Optimal set	79.80	0.107	9862
$(\frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5})$	$(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$	77	0.106	3524
$(\frac{2}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6})$	$(\frac{2}{4}, \frac{1}{4}, \frac{1}{4})$	77	0.106	3511
$(\frac{5}{15}, \frac{4}{15}, \frac{3}{15}, \frac{2}{15}, \frac{1}{15})$	$(\frac{3}{6}, \frac{2}{6}, \frac{1}{6})$	77	0.104	3550

The worst choice in parameters is the one that assumed equal influences at the different historical time points. The highest prediction accuracy was found for the model that uses the parameters deduced from solving the optimization problem. There for relaxing for the parameters being optimal is not considered to be a way to improve the HomcChin2D algorithm although the CPU time is reduced up till 3511 seconds.

Correction with two correctors: the distributions based on the assumption that the first and the third quartile are estimators for the data set on $\mathcal{J}_t$, do give for the PARK1000 a prediction accuracy of 93.94% and a mean range of 0.122. For including the correctors based on the 5th and 95th percentile, the prediction accuracy equals 98,99% and the mean range 0.145. The prediction accuracy increases, and the mean range become less desirable. It is still difficult to predict the scan rate the first 30 minutes of the day.

The adjusted predicted distribution ($w^{(n)}$) resulting from the convex combination that is used to correct for prediction flaw is established now with three distribution vectors: the predicted distribution ($x^{(n)}$) and two other distributions ($y^{(n)}$ and $z^{(n)}$) based on respectively the 25th and the 75th percentile. Such that $w^{(n)}$=$\beta_1 x^{(n)} + \beta_2 y^{(n)} + \beta_3 z^{(n)}$. The default combination uses equal coefficients ($\beta_1 = \beta_2 = \beta_3$ =1/3). Exploration was done on the values of the coefficient used in the convex combination. Some of the results can be found in table 6.2.2b. This table has also the p-value coming from the Shapiro Wilk normality test, used to examine whether it can be said that the a posteriori error come from a normal distribution.

Table 6.2.2b: Prediction accuracy and mean range for scan data by set of coefficients for corrector

$(\beta_1, \beta_2, \beta_3)$	prediction accuracy	mean range	Cpu time seconds	p-value for test normality
Optimal	75	0.093	13001	0.0000
$(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$	93.94	0.122	7369	0.4533
$(\frac{2}{4}, \frac{1}{4}, \frac{1}{4})$	93.94	0.125	5245	0.5784
$(\frac{5}{7}, \frac{1}{7}, \frac{1}{7})$	92	0.130	9419	0.3858

While the optimization with respect to the error to the stationary distribution does not result in improvement of prediction accuracy for the same time periods, the combination $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ and $(\frac{2}{4}, \frac{1}{4}, \frac{1}{4})$ both have the similar prediction accuracy on PARK1000. The mean ranges do not differ too much. And the from the Shapiro Wilk normality test it cannot be concluded that the a posteriori error does not come from a normal distribution (p-value = 0.4533).

Therefore, zooming in on the performances of the prediction accuracy in the n^{th} minute and after *n* minutes was done to decide which model should be given preference. The model with equal coefficients for the correctors had a mean prediction accuracy in the n^{th} minute of 82.73% and in *n* minutes is 83.80%. The mean time that 100% of the data is correctly predicted from minute to minute equals 92.68 minutes. The prediction accuracy is the highest in the first minute. It was tested whether for both types of prediction accuracy equal 95% and more. At significance level 5% the null hypotheses are not rejected.

The data does not convince us that the prediction accuracy in the n^{th} minute and the prediction accuracy after *n* minutes does not exceeds 95% for the model with equal coefficients. In the table 6.2.2c some prediction accuracies can be found

The model with coefficients, $(\frac{2}{4}, \frac{1}{4}, \frac{1}{4})$, for the correctors had a mean prediction accuracy in the n^{th} minute of 83.20% and in *n* minutes is 82.42%. The mean time that 100% of the data is correctly predicted equals 26.52 minutes. For both type of prediction accuracies, it can be concluded that it is not proven that it cannot be more than 90%.

Table 6.2.2c: Prediction accuracy, 100 runs

Minute	n^{th} minute	*n* minutes
1	0.92	1.00
5	0.85	0.92
10	0.85	0.88
15	0.76	0.88
20	0.79	0.87
30	0.85	0.84
200	0.82	0.82

As the performance in minutes of predictions is better can be concluded that the with the equal coefficients has the best performance (CPU=6395s). Noteworthy is that the prediction accuracy in 200 minutes is a time series. Figure 6.2.2 give us some insight in the evolution of the prediction accuracies.

Fig 6.2.2a: Evolution prediction accuracy on PARK1000 in 200 minutes ((1/3 , 1/3, 1/3))

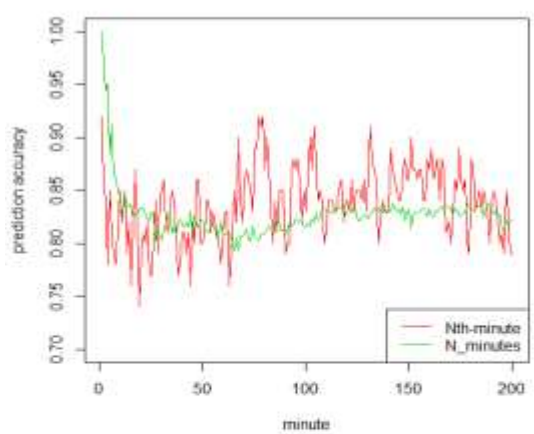
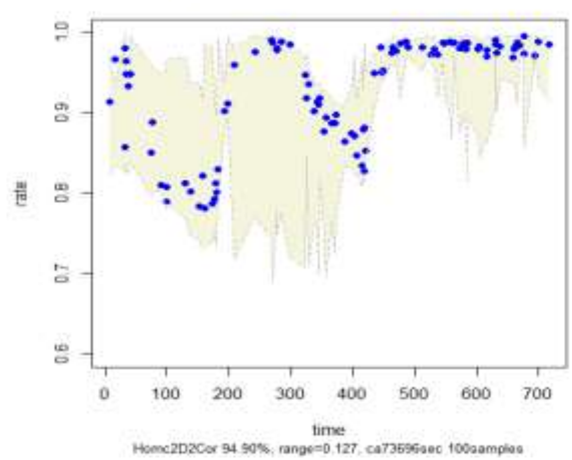


Fig. 6.2.2b Scan rate in an adjusted prediciton interval



Discussion of the research results

This research shows clearly that Markov chain models themselves are not sufficient to do prediction for this time series. Admittedly, first order Markov chain prediction models can imitate the behavior of the time series better than higher order model. But despite this good imitation, the predicted values do underestimate and sometimes even overestimate the real occupation rate. One reason could be that Markov chain prediction models are meant for time independent data. Basically, the same transferring behavior is repeated but with different variants. Markov chain prediction models do predict the distribution after *s* minutes iteratively. Doing this results in an "expectation" for the distribution; a distribution that takes the probabilities of all possible outcomes into account. It turns out to be a probability vector of accumulated probabilities that never occur at time *t*. In real life, not all transition probabilities are "used" to decide the next number of cars at time *t*.

Therefore, additional mathematical model techniques are needed to take care of the aspects of the problem that are beyond the scope of Markov chain prediction models. The actual predictions are done with a Markov chain prediction model. The mathematical techniques help to take care of the presence of two distinct groups in the process. It might be the case that using the point in between where the bimodal distribution of the parking time tails of is too simple and might not work for future data set. Assuming independency between these two groups is disputable. If one chose to assume dependency, a model that takes the interaction between these groups would be needed.

The implementation of the convolution in the algorithm has also side effects. As it is a busy parking place, the probabilities on large numbers of cars for both groups is high. Using at least one distribution with high probabilities for large numbers of cars, to find the convolution, will result in a distribution that might over emphasize the probabilities on large states. Such that the expected value for the prediction will be an occupation rate larger than the real one.

On the other side, the convoluted distribution can lead to predictions for the occupation rate that under estimate the real values. The convoluted distribution was modified, by accumulating all the probabilities that the convolution computed for the number of cars equal to and higher than the maximum number of cars in the system. It could be that to a certain extent the convolution should be associated with the number of cars cruising in the neighborhood looking for an empty space. If it was clear what fraction of the loss probability should be included in the model each time and the day, the state space could be extended, and the convolution could be better modified.

In this research it is thought that more points are needed to predict the next data points. This is especially true in case of large changes in the number of cars between two consecutive minutes. Taking more historical points into account when making predictions will result in a combined transfer to the next time point. This thought is explored by using higher order models. Despite this benefit higher order models lead to prediction intervals with large mean ranges such that large prediction errors are allowed.

Corrections for prediction flaws were done every minute. The ideal situation is of course that the mean range for the predictions reflect the mean range deduced from the data before any prediction is done. Both ways, the mean range for the priori error and the mean range deduced from the prediction interval based upon the historical data way to correct for prediction flaws is to use the distribution of the ranges of the a-priori error, could be met in the markov2D models. The higher order models and correction for prediction flaws both violated this ideal situation such that the ultimate model had a relatively large mean range. Noteworthy is that the mean range of the "best model" can be allowed as it lies somewhere between the minimum and the maximum net added number of cars per minute. Doing predictions with a for example 3 historical point with large transfer in between, would imply working with a larger range.

Conclusion

The company wants to know to what extent predictions could be done for the parking occupancy in a closed neighborhood based on data made available by ARS. A higher order models as proposed by Chin sustained by mathematical techniques was found. This resulted in an algorithm that acknowledges independence behavior of two significant groups at the parking place. For each of this dimension it predicts the distribution of the number of cars on PARK200 with the higher order Markov chain prediction model. This Markov chain prediction model uses the fit distribution of the transitions in a forward moving interval with time span of 15 minutes to estimate the transition probability matrix on time *t*. The distribution for the number of cars on PARK200 is found with the convolution of the predicted distributions for both dimensions of the chain. Corrections for random noise are done for each estimation of the distribution for PARK1000 on time t using a convex combination or equally weighted the distribution of the estimated number of cars on PARK1000 together with an estimated distribution for the 25^{th} and the 75^{th} percentile of the number of cars in the associated time interval.

The mean of all the prediction accuracies in the n^{th} minute equals 82.73%. The mean of all prediction accuracies of all periods of n minutes equals 83.80%. The mean time that 100% of the data is correctly predicted equals 92.68 minutes. In the 15^{th} minute the mean prediction accuracy equal 76% and in the 30^{th} minute 85%. For the 100 samples the prediction accuracy equals 93.94% and its mean range 0.107. At significance level of 5%, the data does not convince us that the prediction accuracy in the n^{th} minute and after n minutes does not exceeds 95% for the model with equal coefficients. At significance level the data did not give enough reasons to conclude that the a posteriori error does not come from a normal distribution. Hence, the model could be used by the company.

In this study the optimum fraction of parking spaces that should be equipped with a sensor was also examined. For the sensitivity of this fraction it could be said briefly: The more with sensor rates equipped columns, the less the error, the better a prediction model. This relation between the number of columns (y*)* and the mean error (*x)* could be approximated the with *y=392.68*exp(-63.62x).* Given the data of the scan vehicles it is even clear that depending on the error the company wants to make, more spaces can be equipped with sensors.

This study reveals that knowledge about the distribution of parking times for visitors and for permit holders is very important. From the study it is clear that there are two groups at the parking place. The impact of these groups on the occupation rate cannot be neglected; not only the parking time but also the number of cars have a bimodally distributed. Based on the bimodal distribution of the number of cars this group is split. A better split would have been possible if it was clear from the data which car was a visitor and which on a permit holder. The latter would be a more exact approach of the group, that could lead for example to better transition probabilities.

The current situation is that the scan vehicles do scan the neighborhood once or twice a day, at a random time point. By being able to predict the scan data this data set and this model does not give us the impression that a change in the number of scans is needed. But as it is clear, the obvious division of the day is a partitioning in eight periods. So, it would be a good idea, to scan the neighborhood at least eight times a day; taking more scans, a day and distributing these scans over the different periods of a day, can give more understanding of the actual error in measurement. It can also help to detect the real a-priori error better. It could also help to generate data in a structured way for further study. Changing the number of the scans per day might help to evaluate the model performance of the model

There are other data sources that can provide information. As parking is not an independent phenomenon, more variables that influence it should be taken into account. Some other factors that influence the parking behavior are the parking demand, the number of available parking places, the parking price. All these factors are time dependent and differ from day to day. There are models developed by other researcher that include these variables in predicting the occupancy rate on a parking place.

My advice to the company is to collect the next set of data systematically for future research. Registering to what group a parker belongs (long- or short-term parker) could leads to a more realistic division between the groups, especially if it regards the last minute of the day. The scan vehicles should start scanning the neighborhood at least eight times a day in the following time periods: 0-30, 30-179, 179-218, 218-313, 313-420, 420-500, 500-719 such that differences within a day can be evaluated. As the error in measurement does not seem to play a big role for finding a model, it would be at least fine to register also the end time of scanning the neighborhood.

The Markov chains lack accuracy when history matters; too much mathematical techniques should be added to the process of such that the prediction is accurate. Therefore, it would be good to use some other models. One such a group of models is the general linear mixed modelling, statistical model for multivariate data. These models can take the time as main predictor variable and while taking every other thinkable variable together with the interactions between these variables simultaneously into account.

As the net added number of cars each minute is a mean zero process (section 1.3.5), it can be evaluated whether Brownian motions could be used to find a more straight forward prediction model. Then come the datamining with its deep learning based on different mathematical techniques such as the nearest neighbor that can be used to keep the data clustered when doing predictions.

Appendix

As it is my wish to help simple people to understand the research it is chosen to put some definitions and derivations needed for the main mathematical concepts used in this report in the appendix.

Appendix 1 Ranges Prediction Interval

In this part of the appendix the ranges for a prediction interval of the occupation rate on PARK1000 at time *t* will be derived using the data from PARK200. The occupation rate equals the fraction of occupied parking spaces on associated time point. As, the sensored spaces are a part of the total parking place it is assumed that the fractions of occupied spaces for both groups are similar. In order to find the ranges for the prediction interval it is good to zoom in on the fundamental processes. For each parking space there are two options: occupied or not. Independent from what happens in other parking space, a parking spaces can be occupied or not at time *t* with fixed probability. The occupation status of each parking space is identically and independent distributed random variable that is Bernoulli distributed with probability *p(t).* A reader must keep in mind that all the variables do depend on the actual time point *t* of the process.

Suppose:

$$Y_i^{(t)} = \begin{cases} 1 \text{ if the } i^{th} \text{ scanned parking space is occupied on time } t \\ 0 \quad \text{else} \end{cases},$$

$Y_i^{(t)} \sim Bern(p(t))$, Y_i independent identically distributed random variable, *i=0, 1, . . . ,m,* where *m* is the total number of spaces at PARK1000.

$$X_i^{(t)} = \begin{cases} 1 \; if \; the \; i^{th} \text{ sensored parking space is occupied on time } t \\ 0 \quad \text{else} \end{cases},$$

$X_i^{(t)} \sim Bern(p(t))$, $X_i^{(t)}$ independent identically distributed random variable, *i=0, 1, . . . ,n,* where *n* is the total number of spaces at PARK200.

Denote

The number occupied parking spaces on PARK1000 is represented by random variable $Y^{(t)}$, with $Y^{(t)}=\sum_{i=1}^{m} Y_i^{(t)}$, and $Y^{(t)} \sim bin(m, p(t))$, where *m*=1000 equals the total number of occupied parking places and *p(t)* the fraction of parking spaces at time *t* (*t*=0, 1, . . . , 719).

The number occupied parking spaces on PARK200 is represented by random variable $X^{(t)}$, with $X^{(t)}=\sum_{i=1}^{m} X_i^{(t)}$, $X^{(t)} \sim bin(n, p(t))$, where *n*=200 equals the total number of occupied parking places and *p(t)* the fraction of parking spaces at time *t* (*t*=0, 1, . . . , 719).

As *n* and m are large the distributions of $Y^{(t)}$ and $X^{(t)}$ can be approximated by the normal distribution.

$$Y^{(t)} \sim bin(m, p(t)) \approx N\big(m * p(t), m * p(t) * (1 - p(t)\big)$$
$$X^{(t)} \sim bin(n, p(t)) \approx N\big(n * p(t), n * p(t) * (1 - p(t)\big)$$

Then the fraction of occupied parking spaces on PARK200, $\hat{p}(t) = \frac{\sum_{i=i}^{n} X_i^{(t)}}{n}$, follows a normal distribution with $\mu = \hat{p}(t)$ and $\sigma^2 = \frac{\hat{p}(t)*(1-\hat{p}(t))}{n}$. $\left(\hat{p} \sim N\left(\hat{p}(t), \frac{\hat{p}(t)*(1-\hat{p}(t)}{n}\right).\right)$

$\hat{p}(t)$is a good predictor for, $\frac{Y^{(t)}}{m}$, the fraction of occupied parking spaces at PARK1000.
$Y^{(t)}$, the number of occupied parking spaces at PARK1000, can be predicted by $m\hat{p}(t)$ with $m\hat{p}(t) \sim N\left(m\hat{p}(t), m^{2*}\frac{\hat{p}(t)*(1-\hat{p}(t)}{n}\right)$.

The random variable $Y^{(t)} - m\hat{p}(t) \sim N\left(0, m * p(t) * (1 - \hat{p}(t) + \frac{m^2}{n}\hat{p}(t) * (1 - \hat{p}(t))\right)$ and as the population variance can be estimated by the variance of the known data the variable
$Y^{(t)} - m\hat{p}(\mathrm{t}) \sim t_{\frac{1}{2}\alpha,(n-k+1)}\left(0, m * \hat{p}(t) * (1 - \hat{p}(t) + \frac{m^2}{n} * \hat{p}(t) * (1 - \hat{p}(t))\right)$
Where $\hat{p}$(t) is a good estimator for the mean fraction of occupied sensored spaces and k=2, the number of involved random variables. The borders for a prediction interval for $Y^{(t)}$ would be:

$$m\hat{p}(t) \pm t_{\frac{1}{2}\alpha,(n-k+1)}\sqrt{m\hat{p}(t)(1-\hat{p}(t)) + \frac{m^2}{n}\hat{p}(t)(1-\hat{p}(t))},$$

The borders of the prediction interval for the occupation distribution of the rate at each time point would be:

$$\hat{p}(t) \pm t_{\frac{1}{2}\alpha,(n-k+1)}\sqrt{\frac{\hat{p}(t)(1-\hat{p}(t))}{m^2} + \frac{\hat{p}(t)(1-\hat{p}(t))}{n}},$$

Appendix 2 Convolution and Strong Law Large Number

Appendix 2 contains two main concepts this study heavily relies on: the convolution and strong law large numbers. Hence the definitions can be found here in case the reader is not acquainted with these concepts.

Definition 1: The convolution of two probability mass functions

Let N_S and N_L be two independent integer-valued random variables, with probability mass function $x_{N_S}(j)$ and $x_{N_L}(j)$ respectively. Then the convolution of $x_{N_S}(j)$ and $x_{N_L}(j)$ is the probability mass function

$$x_N = x_{N_L} * x_{N_S} \text{ given by}$$

$$x_N(j) = \sum_{k=0}^{M} x_{N_S}(k) * x_{N_L}(j-k) \text{ , for } j = 0, 1, 2, \ldots M$$

The function $x_N(j)$ is the probability mass function of the random variable N, where *n* represents the number of cars at an arbitrary time point with $N = N_S + N_L$. *(Grinstead et al, 1994, page 286, Mandal P.K., 2017, p.24, 25)*

Theorem 1: Strong law of large numbers *(Weisstein, 2019)*

The sequence of variates X_i with corresponding means μ_i obeys the strong law of large numbers if $\exists\ N, \exists \varepsilon, \delta > 0$ that for every r>0, all (r+1) inequalities, $\frac{|S_n - \hat{S}_n|}{n} < \varepsilon$, for n=N, N+1, . . . , N+r will be satisfied where $S_n = \sum_{i=1}^{n} X_i$, and $\hat{S}_n = \sum_{i=1}^{n} \mu_i$.

Appendix 3 Distribution of the Sum of Independent Poisson Variables

For a Poison process with arrival rate λ it is generally known that the number of events in any interval of length *t* is Poisson distributed with mean *λt*. For the Poison process with arrival rate λ at the parking place it can similarly be said that the number of events in any parking space *n*, is Poisson distributed with mean *λn.* Here to Lemma 1 is proven:

Lemma 1:

Given s independent Poisson random variables, $X_i \sim Pois(\alpha_i)$ *for i =1, 2, …., and Z=*$\sum_{i=1}^{s} X_i$*, then random variable Z is Poisson distributed with mean E(Z)=* $\sum_{i=1}^{s} \alpha_i$, $\forall s \in \mathbb{Z}^+$.

Proof:

We will proof by induction that for *s* independent Poisson random variables, $X_i \sim Pois(\alpha_i)$ for *i* =2, ., *s* the random variable Z=$\sum_{i=1}^{s} X_i$ is Poisson distributed with mean $\sum_{i=1}^{s} \alpha_i$.

Proof:

The proof is done by induction.

Base case: For *s*=2 we have:

$X_1 \sim Pois(\alpha_1)$ and $X_2 \sim Pois(\alpha_2)$, $Y = X_1 + X_2$,

$$P(Y=y) = P(X_1 + X_2 = y) = \sum_{x=0}^{y} P(X_1 = x, X_2 = y - x) = \sum_{x=0}^{y} P(X_1 = x)P(X_2 = y - x).$$

Substituting the Poisson distribution and then multiplying with $\frac{y!}{y!}$gives: $\sum_{x=0}^{y} e^{-\alpha_1} \frac{{\alpha_1}^x}{x!} e^{-\alpha_2} \frac{{\alpha_2}^{(y-x)}}{(y-x)!} \frac{y!}{y!}$

$$P(Y=y) = \frac{e^{-(\alpha_1+\alpha_2)}}{y!} \sum_{x=0}^{y} \binom{y}{x} {\alpha_1}^x {\alpha_2}^{y-x} = \frac{e^{-(\alpha_1+\alpha_2)}}{y!} (\alpha_1 + \alpha_2)^y = e^{-(\alpha_1+\alpha_2)} \frac{(\alpha_1+\alpha_2)^y}{y!}.$$

At this point it can be concluded that $Y \sim Pois(\alpha_1 + \alpha_2)$.

Induction step: Let $n \in \mathbb{Z}^+$ be given and suppose Lemma 1 is true for *s=n+1*.

$$P(Y=y) = P\left(\sum_{i=1}^{n+1} X_i = y\right) = P\left(\sum_{i=1}^{n} X_i + X_{n+1} = y\right) =$$

$$\sum_{x=0}^{y} P\left(\sum_{i=1}^{n} X_i = x, X_{n+1} = y - x\right) = \sum_{x=0}^{y} P\left(\sum_{i=1}^{n} X_i = x\right) P(X_{n+1} = y - x) =$$

$$\sum_{x=0}^{y} e^{-(\sum_{i=1}^{n} \alpha_i)} \frac{(\sum_{i=1}^{n} \alpha_i)^x}{x!} e^{-\alpha_{n+1}} \frac{{\alpha_{n+1}}^{(y-x)}}{(y-x)!} = \sum_{x=0}^{y} e^{-(\sum_{i=1}^{n} \alpha_i)} \frac{(\sum_{i=1}^{n} \alpha_i)^x}{x!} e^{-\alpha_{n+1}} \frac{{\alpha_{n+1}}^{(y-x)}}{(y-x)!} \frac{y!}{y!} =$$

$$\frac{e^{-(\sum_{i=1}^{n} \alpha_i + \alpha_{n+1})}}{y!} \sum_{x=0}^{y} \binom{y}{x} \left(\sum_{i=1}^{n} \alpha_i\right)^x {\alpha_{n+1}}^{y-x} \frac{e^{-\left(\sum_{i=1}^{n+1} \alpha_i\right)}}{y!} \left(\sum_{i=1}^{n} \alpha_i + \alpha_{n+1}\right)^y \frac{e^{-(\sum_{i=1}^{n+1} \alpha_i)}}{y!} \left(\sum_{i=1}^{n+1} \alpha_i\right)^y.$$

Thus, Lemma (1) holds for s = n + 1, and the proof of the induction step is complete. The conclusion is that Lemma (1) is true $\forall s \in \mathbb{Z}^+$.

Appendix 4 Words, concepts and abbreviations

ARS T & TT	ARS Traffic & Transport Technology. The first part of the name ARS comes from the Greek (Latin) word “ars” which was related to craft, skill, knowledge, method, device, or even science.
CBS	Centraal Bureau voor de Statistiek
CPU time	Central Processing Unit time. The amount of time used to process the instructions in a central processing unit of a computer program
Cruising	Looking for a parking place while driving. Sometimes even literally circling in the neighborhood of the parking place. Cars that are cruising are also referred to as search traffic
Distribution of scans over the day	The collection of time periods that the scan vehicle passed in the neighborhood during the day
Estimated rate	The estimated percentage of the occupied parking spaces
HomcChin2D	The algorithm in this report used for doing predictions with for a two-dimensional Markov chain. The actual predictions are done with the higher order Markov chain model as proposed by Chin et. al..
Kendall's correlation	A nonparametric measure of the strength and direction of association that exists between two variables measured on continuous or ordinal scale. It is considered a nonparametric alternative to the Pearson’s product-moment correlation. A monotonic relationship if desirable but not a strict assumption *(Laerd Statistics, 2018).*
Long-term parkers	Users of the parking place with a parking duration longer than 630 minutes (10.5 hours). Usually the residents and / or permit holders.
Markov2D	The algorithm in this report used for doing predictions with for a two-dimensional Markov chain. The actual predictions are done with the first order Markov chain model.
Occupation rate	The percentage of the parking spaces that are occupied
PARK1000	The whole parking place in the closed neighborhood consisting of 1000 parking spaces.
PARK200	The part of the parking place that is simulated as if each of these arbitrarily chosen 200 spaces were “equipped” with sensors.
Parking place	In this research it is the set of all the public parking spaces along the street or streets in a neighborhood
Parking space	A square in the parking place intended for parking exactly one vehicle.
Predicted rate	The predicted percentage of the occupied parking spaces
Scan-rate	The rate computed with the scanned data
Search traffic	See Cruising

Sensitivity to the distribution of scans	The extent to which a change in the distribution of the number of scans per day influences the end results.
Sensitivity to the fraction	The extent to which a change in the fraction influences the end results.
Sensitivity to the number of scans per day	The extent to which a change in the number of scans per day influences the end results.
Sensor	A small smart device mounted in a parking space that gives an infra-red signal to an intelligent network system to indicate when the parking space is occupied.
Sensor rate	The percentage of the occupied parking spaces computed with the sensor data, or the data from PARK200
Short-term parkers	Users of the parking place with a parking duration not longer than 630 minutes (10.5 hours). In this report the are also identified as “visitors”.
Smirnov test	A nonparametric test that is used to decide whether two samples do have a similar continuous distribution. The test statistic is based on the cumulative distribution at the point of the maximum deviation between the associated cumulative distributions.

References

ARS T&TT, 2015, *About ARS T&TT*, https://ars-traffic.com/en/about-ars-ttt , http://www.techniekwerkt.nl/bedrijf/ars-t-tt-den-haag-4377

BachMaier S, 2016, *Performance Evaluation of Time-Based and Movement-Based Location Update Schemes*, https://elib.uni-stuttgart.de/bitstream/11682/2640/1/thesis.pdf

Belloche, S., 2015, On-street Parking Search Time Modelling and Validation with Survey based Data. *Transportation Research Procedia*, http://doi.org/10.1016/j.trpro.2015.03.024

Bolle, E.A.W, Lenoir, J.H.M., van Loon, J.N.M, 1974, *Wiskundige statistiek*,Kluwerbv, Deventer

Bork van G.J., Struik H., Verkruijsse P.J., Vis G.J. ,2002 *, "ARS" in Letterkundig lexicon voor de neerlandistiek*, https://dbnl.org/tekst/bork001lett01_01/bork001lett01_01_0002.php#a199

Breunesse, E., 2018. '*Shell and the Energy Transition (in The Netherlands) general'*, Vexpan Nationaal Parkeercongres 2018, Eindhoven

CBS , 2017, *Bevolkingstrends-Nederlanders en hun auto - Een overzicht van de afgelopen tien jaar*, https://www.cbs.nl/nl-nl/achtergrond/2017/08/nederlanders-en-hun-auto

CBS, 2018, *Grootste groei toerisme in ruim tien jaar*, https://www.cbs.nl/nl-nl/nieuws/2018/14/grootste-groei-toerisme-in-ruim-tien-jaar

Ching W, Huang X, Ng M.K., Siu T.K., 2013, *Higher-Order Markov Chains. In Markov Chains,* Springer

Ching W, Ng MK, Fung E.S., 2008, *Higher-order multivariate Markov chains and their applications. Linear Algebra and its Applications*

Ching W., Ng M.K., 2006, *Markov Chains: Models, Algorithms and Applications*

Coghlan A., 2010, *Little Book of R for Time Series!*, a https://a-little-book-of-r-for-time-series.readthedocs.io/en/latest/

CROW Ede, 2014, *Trends in parkeren*, http://www.crow.nl/getmedia/94d92f23-728f-4e1e-b374-246039a2fe59/Trends-in-parkeren-def-klein-v1-

Gantelet, E., & Fauconnier, A. le., 2006. *The time looking for a parking space: strategies, associated nuisances and stakes of parking management in France.* In European Transport Conference (ETC), http://trid.trb.org/view.aspx?id=846283

Grinstead C, Snell L, 1994, *Introduction to probability*, https://www.dartmouth.edu/~chance/teaching_aids/books_articles/probability_book/Chapter7.pdf

Hurink J.L., 2017, *Heuristic Methods in OR,* slides University of Twente

INRIX Research, Graham Cookson, 2018, *INRIX Global Traffic Scorecard*

Laerd Statistics, 2018, *Pearson's Product-Moment Correlation using SPSS Statistics,* Lund Research Ltd, https://statistics.laerd.com/spss-tutorials/pearsons-product-moment-correlation-using-spss-statistics.php

Leiden, 2014, *Rapportage Parkeermonitor Leiden Juli - december 2013*, Versie 1.1

Liu Tie, April 2010, *Application of Markov Chains to Analyze and Predict the Time Series*

Mandal P.K., 2017, *Introduction to Risk Theory, Quarter 2, 2017-18,* slides University of Twente

Meijer, C., 2018. *'Parkeren in de mobiele toekomst'*, Vexpan Nationaal Parkeercongres 2018, Eindhoven

OIS Amsterdam, 2017, *Amsterdam in cijfers*, Jaarboek 2017

Raftery A, 1985, A Model for High-order Markov Chains, *Journal of Royal Statistical Society*, Series B,

Ross, S. M., 2010, *Introduction to probability models.*,10th ed

Rumsey D.,J., , 2018, *What the Distribution Tells You About a Statistical Data Set*, https://www.dummies.com/education/math/statistics/what-the-distribution-tells-you-about-a-statistical-data-set/

Shoup D., 2005, *The High Cost of Free Parking's, Park-and-visit test in Westwood village*

Shoup, D., 2005b, The high cost of free parking. Planners Press, *American Planning Association*

Shoup, D., 2006, *Cruising for parking*. Transport Policy, http://doi.org/10.1016/j.tranpol.2006.05.005

Smetana dr. K., 2018, *"Scientific Computing"* Lecture notes University of Twente

Stienstra Ir. Sj. Adviesbureau stedelijk verkeer BV, 2011, *Nederlands Parkeerbeleid in Europees Perspectief*, https://www.cvs-congres.nl/cvspdfdocs/cvs11_007.pdf

Tiedemann, T., Vögele, T., Krell, M.M., Metzen, J.H., Kirchner, F., 2015, *Concept of a data thread based parking space occupancy prediction in a berlin pilot region*. In: Papers from the 2015 AAAI Workshop (WAIT-2015). AAAI Press
Van Ommeren, J., Wentink, D., & Rietveld, P. (2012). *Empirical evidence on cruising for parking. Transportation Research Part A: Policy and Practice*, http://doi.org/10.1016/j.tra.2011.09.011
Vlahogianni*E., I., Ph.D., Kepaptsoglou, K., Ph.D., Tsetsos V., Ph.D. and Karlaftis, M., G., Ph.D. , 2015, *A Real-Time Parking Prediction System for Smart Cities, Journal of Intelligent Transportation Systems*, DOI: 10.1080/15472450.2015.1037955
Yadav D, Kang T.S., Spedicato G.,A. *Higher, possibly multivariate, Order Markov Chains in markovchain package*, https://cran.r-project.org/web/packages/markovchain/vignettes/higher_order_Markov_chains.pdf
Weisstein, E. W., 2019-02-04, "*Strong Law of Large Numbers",* From MathWorld--A Wolfram Web Resource, http://mathworld.wolfram.com/StrongLawofLargeNumbers.html
Wu T., Gleich D., 2017, *Retrospective Higher-Order Markov Processes for User Trails.*