Simulating Formula One Race Strategies

Claudia Sulsters (2540240)
Research Paper MSc Business Analytics

Vrije Universiteit Amsterdam
Faculty of Sciences
De Boelelaan 1081a
1081 HV Amsterdam

Supervisor:
R. Bekker

February 26, 2018

Abstract

The goal of this research is to build a simulation model that can be used by Formula One teams to determine the optimal race strategy among a set of possible strategies. The simulation model describes the influence of fuel consumption and tire degradation on the lap times based on observed lap time data from races of the 2016 season. In addition, the simulation model imitates most on-track events, including the mixing of the cars at the start of a race, pit stops, overtaking actions, safety car situations and driver retirements. The performance of the simulation model is evaluated using four races of the 2016 season. The results show that the simulated positions are highly correlated with the actual positions, implying that the model performs well in predicting the race results. However, the race times are often under- or overestimated. Finally, we illustrate how the simulation model can be used by Formula One teams by simulating different race strategies and comparing the results.

Keywords: Formula One, Discrete Event Simulation, Linear Regression, Bayesian Inference.

Preface

The amounts of data that are generated in Formula One inspire me in my work as a mathematician, but also the sport itself intrigues me. After my first actual visit to a Formula One race, the 2017 Spanish Grand Prix, I have never missed a free practice, a qualification session or a race. When I saw the fast cars driving around the corners of the circuit de Catalunya, I knew that I wanted to learn more about the sport. Therefore, it was not difficult to choose the subject of my research paper for the master Business Analytics. This research paper aims at simulating race strategies for Formula One races using open-source data. It combines my passions for Formula One, mathematics and data science.

I would like to thank my supervisor René Bekker for the advice and guidance throughout the research process.

Contents

1 Introduction

One may say that Formula One races are won in the factory instead of on the circuit. Formula One teams collect tremendous amounts of real-time data in order to predict the performance of their race cars. The McLaren team, for example, builds its race cars for each circuit based on historical data and simulations generated by the current season's sensor data. The Red Bull Racing team has sixty engineers present in the pit box and thirty in England during a Formula One race weekend. The data from the farthest circuit in Australia reaches the Red Bull's U.K. team in less than 300 milliseconds, to run simulations to determine or adjust the race strategy whenever a tire change or an overtaking opportunity occurs [1]. This is because strategy decisions can be crucial in winning a Formula One race. Simulating the result of a race given a particular strategy can help Formula One teams planning and evaluating their race strategies, possibly leading to competitive advantages.

J. Bekker and W. Lotz (2009) have designed and evaluated a discrete-event simulation model that can be used for planning Formula One race strategies. Bekker and Lotz divide the circuit in a number of sectors and simulate a sector time for each sector using detailed on-track data of Formula One cars such as fuel consumption, tire degradation and air resistance. The simulation model imitates most on-track events, including car failures, overtaking, and pit stops [2].
However, the simulation model was published by Bekker and Lotz back in 2009 and cannot be used to simulate Formula One races today, because of the regulations being yearly revised by the Fédération Internationale de l'Automobile's (FIA), the governing body of Formula One. As a result, the regulations during the 2009 season differ strongly from the regulations of the 2016 season. First, the FIA has decided that refueling during the race is no longer permitted from the 2010 season onwards. This implies that refueling is no longer included in a pit stop, while it was in 2009. Second, the Drag Reduction System (DRS), which facilitates overtaking between Formula One cars, was introduced in 2011. As a result, the average number of overtaking actions per race has increased from 28.79 in 2010 to 60.63 in 2011 [3]. Since the temporary speed advantage resulting from the DRS is not taken into account by Bekker and Lotz, the overtaking model needs to be adjusted in order to be used in a simulation model today. Third, the deployment of the safety car, which can highly influence the race strategy, is not included in the model of Bekker and Lotz. Finally, it should be noted that Formula One cars are improved each season to be able to drive faster and to make races more spectacular, which makes the data on the performance of race cars in 2009 unusable today. In this research, a discrete-event simulation model is developed that is able to simulate races of the 2016 season. Since the model uses lap time data from races of the 2016 season, the performance of the 2016 Formula One racing cars is very well reflected in the model. Furthermore, most important changes in regulations, such as the use of the DRS, and the deployment of the safety car are included in the model.

The research goal is to build a simulation model that can be used by Formula One teams to determine the optimal race strategy among a set of possible strategies. We define the optimal race strategy as the strategy that maximizes the expected positions of both drivers in the team.
The simulation model describes the influence of fuel consumption and tire degradation on the lap time based on observed lap time data from races of the 2016 season. In addition, the simulation model imitates most on-track events, including the mixing of the cars at the start of a race, pit stops, overtaking actions, safety car situations, and driver retirements. This research is limited to the use of open-source data that can be found on the internet, since Formula One racing teams do not have their data freely available.

This paper is structured as follows. Section 2 provides the reader with relevant background information on the subject of Formula One racing. The data that is used in this research is analyzed in Section 3. Section 4 focuses on the design and the implementation of the simulation model. Finally, the results are presented and discussed in Section 5, whereafter conclusions are drawn in Section 6.

2 Formula One background

Formula One originates from the pre-war European Grand Prix championships of the 1920s and 1930s. However, we may say that it was officially founded in 1946, when the FIA introduced the first standardized rules. In 1950, the FIA introduced the first official World Championship for Drivers using the Formula One rules. This World Championship consisted of six European Grand Prix and the Indianapolis 500. 'Formula One' refers to a set of technical regulations for single-seater open-cockpit racing cars. These regulations are published annually by the FIA.

The 2016 FIA Formula One World Championship consisted of twenty-one Grand Prix located all over the world. Eleven teams (constructors) consisting of two drivers competed in two championships: the Constructor's World Championship and the Drivers' World Championship. The driver that has the most championship points receives the Driver's World Championship title, while the teams compete for the Constructor's World Championship title. The drivers and teams are awarded championship points each Grand Prix based on their final positions.

A race weekend consists of two free practices on Friday, a free practice and a qualifying session on Saturday, and the race on Sunday. During the free practice sessions, the circuit is available for the teams and the drivers to work on the set-up of their cars in preparation of the qualifying session and the race. Engineers and mechanics use detailed on-track data of the racing cars to optimize the performance of the car on the circuit. In general, a balance must be found between the top speed of a Formula One car and its down force. High down force is necessary to drive fast through the corners of the circuit, while a high top speed is necessary on the long straights. Not having the right balance in the car will cause understeer or oversteer in the corners. Understeer is used to refer to the situation that the front end of the car refuses to turn into a corner and slides wide, while oversteer refers to the situation that the rear end of the car refuses to go around the corner and tries to overtake the front end of the car. Furthermore, the aerodynamica has a high influence on the performance of a Formula One car. The chassis, the main part of a racing car to which the engine and suspension are attached, is therefore also optimized by the engineers.

On Saturday, the drivers compete in the qualifying session trying to achieve a high position on the starting grid. The qualifying format was revised after the first two races of the 2016 season. Here we discuss the format that was used for the majority of the races. In this format, the qualifying session consists of three knock-out sessions: Q1, Q2, and Q3. The first knock-out session, Q1, has a duration of eighteen minutes in which all cars will be permitted on the circuit. At the end of the session, the slowest six cars are excluded from the qualifying session and the lap times achieved by the sixteen remaining cars will then be deleted. The second knock-out session, Q2, has a duration of fifteen minutes. The sixteen remaining cars will be permitted on the track and at the end of this period again the slowest six cars are excluded from the qualifying session, while the lap times achieved by the ten remaining cars will be deleted. The last session, Q3 or top 10 qualifying, has a duration of twelve minutes. The remaining ten cars are permitted on the circuit to compete for the top 10 positions on the starting grid. Especially, the battle for pole position, the first place on the starting grid, is a fierce one. Although the qualifying sessions determine the position of the drivers on the starting grid, drivers can receive a grid penalty of a few positions by violating the technical regulations of the FIA.

The Formula One teams and drivers compete for the championship points during the race on Sunday. The most spectacular moment is probably the start of the race, since this is one of the best opportunities to gain one or more positions. The race starts when the red lights go out, and the field accelerates away towards the first corner. It is not unusual to see two or more cars simultaneously taking the first corner. Contact between cars is sometimes unavoidable, as the cars are heavy with fuel, have relatively cold brakes and tires and are off the normal racing line trying to overtake each other.

During the race, each driver may visit his team in the pits to replace his current set of tires by a fresh set of tires, which is known as a pit stop. The tire producer of Formula One tires, Pirelli, produces five different types of dry-weather tires: ultrasoft tires, supersoft tires, soft tires, medium tires, and hard tires. These different types of tires differ in durability and grip. In general, softer compounds have more grip and lower durability, while harder compound have less grip but high durability. In addition, Pirelli produces two wet-weather tires, which are the intermediate tires and the wet tires. The characteristics of the different tire compounds are summarized in Table 1. Each car has thirteen sets of dry-weather tires, four sets of intermediate tires and three sets of wet tires available during a Formula One race weekend. The Formula One teams have access to three different compounds of the dry-weather tires. Pirelli nominates two mandatory sets for each car for the race (which can be of different compounds) and one set of the softest compound that can only be used in the Q3 qualifying session. Usually, Pirelli chooses a 'prime tire' and an 'option tire'. The prime tire is in theory most appropriate for the circuit's characteristics and is normally harder than the option tire. The option tire is not expected to be as appropriate as the prime tire, but may provide certain advantages in terms of pace or durability. The Formula One teams are free to choose the remaining ten sets of dry weather tires.

A decent race strategy is essential for winning races. A race strategy is determined by the choice of the tires and the laps in which the pit stops are made. One important aspect of planning a pit stop is where in the field the driver would reemerge after the pit stop. One well-known pit stop strategy is 'undercutting' or 'overtaking in the pit lane'. Assume that a car is right behind a slower car that is hard to overtake. The trailing car can decide to make an early pit stop, then return on the track in clean air, and drive some fast laps on his new tires. Using this strategy, the faster car tries to ensure that the slower car emerges behind him after the slower car makes his pit stop. Since it is not known in advance of the race whether such an overtaking opportunity will occur, Formula One teams use a system of pit stop windows instead of fixed pit stop timetables. Furthermore, it can be very beneficial to make a pit stop when the safety car is deployed. A safety car can be deployed by the Race Director when he wants to reduce speed for safety reasons, for instance, after an accident or because of heavy rain.
After the race, the top three drivers are honored in a podium ceremony. The trophies are raised and the championship points are awarded.

Compound	**Driving conditions**	**Grip**	**Durability**
Ultrasoft	Dry	1	5
Supersoft	Dry	2	4
Soft	Dry	3	3
Medium	Dry	4	2
Hard	Dry	5	1
Intermediate	Wet (light standing water)	-	-
Wet	Wet (heavy standing water)	-	-

Table 1: Characteristics of the tire compounds available during the 2016 season. Tires are ranked on grip and durability at a scale from 1 (highest) to 5 (lowest).

3 Data analysis

The data that is used in this research consists of driver's lap times from each race during the 2016 season. This data is retrieved from the Ergast Developer API (`ergast.com`), which is an experimental web service that provides data for the Formula One series from the beginning of the World Championship in 1950 [4]. The 2016 World Championship consisted of twenty-one races all over the world. In the following sections, we will use this data set to illustrate the methods that are used to model the different components of the simulation model. This section analyses the lap time data as retrieved from the Ergast Developer API.

Figure 2 shows Nico Rosberg's lap times for each lap during the Grand Prix of Europe (Baku) in 2016. We can observe that the lap time of the first lap is considerably higher than the lap times that follow the first lap (laps 2-20). This is caused by the start of the race for three important reasons. First, Formula One drivers need tenths of seconds to react to the red lights going out and a few seconds to accelerate away from the grid towards the first corner. Second, drivers try to gain one or more positions in their run to the first corner, resulting in some time loss in the first lap. Third, drivers need to warm up their brakes and tires in the first lap, which again has a negative influence on the lap time.

In addition, we can observe an in-lap at lap 21 and a pit stop in lap 22. During the in-lap, drivers slow down to enter the pit lane entrance, whereafter the driver stops at his pit box to replace his current tires by a new set of tires. During the European Grand Prix, Nico Rosberg completed two different stints (i.e. a set of consecutive laps) on two different sets of tires. However, from the lap time data it is not possible to derive the tire compounds that were used during these stints.

We can also observe a decreasing trend in the lap times over the duration of the race, which is caused by fuel consumption. On the other hand, it is known that the lap times increase at the beginning and the end of a stint. This is caused by a relatively low temperature of the tires at the beginning of the stint and by tire degradation at the end of a stint. We can conclude that we have to estimate the influence of both fuel consumption and tire degradation on the lap times in order to simulate realistic lap times. Finally, some lap time variability is present in the lap times, especially during the second stint.

The original data, as retrieved from `ergast.com`, contains the lap time and position of a driver during each lap of a Formula One race, but does not contain the tire compound that is used in each lap, the age of the tires, and the fuel level. Therefore, this data is enriched with pit stop strategy data and fuel level data to obtain a complete dataset that can be used to estimate the influence of fuel consumption and tire degradation on the lap times. Table 2 shows an example of the enriched dataset for Nico Rosberg during the European Grand Prix in 2016.

From the pit stop strategy data, we can derive that the pit stop of Nico Rosberg in lap 21 had a duration of 20.058 seconds and that his set of supersoft tires was replaced by a set of soft tires [5]. Nico Rosberg's race strategy thus consisted of a stint on supersoft tires (laps 1-21) and a stint on soft tires (laps 22-51). We can now use the age of a set of tires to estimate the influence of the tire degradation on the lap time. However, due to the lack of data, we have to make some assumptions on the fuel consumption. We assume that the fuel level decreases linearly over the duration of the race, having a level of 100% at the start of the race and a level of 0% at the end of the final lap. This implies the assumption that the fuel consumption rate is constant during the race.

Other data resources that are used in this research are qualifying results [6], the starting grid [7] and race results [8] for each race during the 2016 season. Also, we use manually collected data about safety car situations that have occurred during the 2016 season. These data on safety car occurrences can be found in Appendix I.

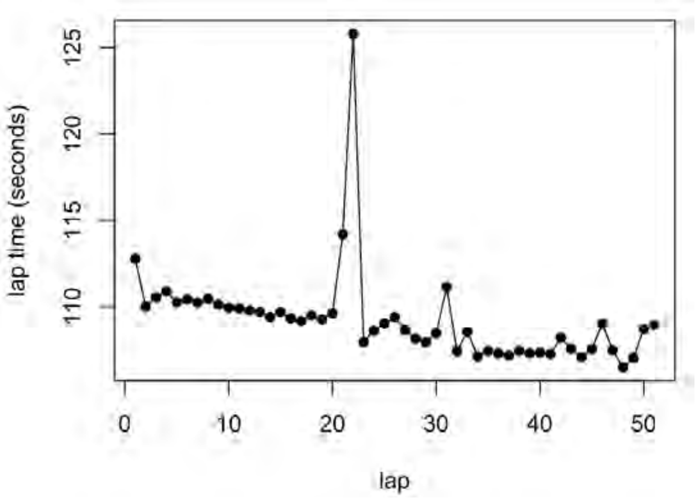


Figure 1: Nico Rosberg's lap times (in seconds) during the European Grand Prix (2016).

Driver	Race	Lap	Position	Lap time	Compound	Tire age	Fuel level %
Nico Rosberg	Europe	1	1	112.772	Supersoft	1	98.039
Nico Rosberg	Europe	2	1	110.007	Supersoft	2	96.078
Nico Rosberg	Europe	3	1	110.541	Supersoft	3	94.118
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Nico Rosberg	Europe	21	1	114.200	Supersoft	21	58.824
Nico Rosberg	Europe	22	1	125.801	Soft	1	56.863
Nico Rosberg	Europe	23	1	107.954	Soft	2	54.902
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Nico Rosberg	Europe	50	1	108.711	Soft	29	1.961
Nico Rosberg	Europe	51	1	108.942	Soft	30	0.000

Table 2: Nico Rosberg's enriched dataset for the European Grand Prix (2016).

4 Model description

The model that is developed in this research uses discrete-event simulation to simulate lap times of drivers during a Formula One race, and, in addition, imitates most on-track events including the mixing of the cars at the start of a race, pit stops, overtaking actions, safety car situations and driver retirements. This model can be used to simulate the results of races of the 2016 season.

The last five races of the 2016 season are assigned to a test set, which can be used to evaluate the performance of the simulation model. However, we have to remove the Brazilian Grand Prix from the test set since the rainy weather conditions during this race violate the model assumptions, which are stated at the end of Section 4.1. The remaining four races are simulated by the simulation model using data from the training set. The training set differs for each of the four races, since it consists of all previous races of the 2016 season. However, the Hungarian Grand Prix is removed from the training set since the qualifying times of the drivers are not representative because of rainy weather conditions during the qualifying session. The parameters for simulating the Japanese Grand Prix, for example, are estimated using lap time data of the first fifteen races. Table 3 shows the training set of the Japanese Grand Prix.

Train	Australia, Bahrein, China, Russia, Spain, Monaco, Canada, Europe, Austria Great Britain, Germany, Belgium, Italy, Singapore, Malaysia
Test	Japan

Table 3: Training set used to estimate parameters for the Japanese Grand Prix.

This section is structured as follows. Section 4.1 discusses the design of the simulation model and presents the model assumptions. Section 4.2 presents the model that describes the influence of the fuel consumption and tire degradation on the lap times. How the different on-track events are included in the model is described in Sections 4.3, 4.4, 4.5, and 4.6. Finally, the methods that are used to evaluate the results of the simulation model are discussed in Section 4.7.

4.1 Model design

This section describes the design of the simulation model and explains how the different components are included in the model. We end this section by summarizing the model assumptions.
According to Law and Kelton (2000) a discrete-event simulation model describes a *system* whose *state* only changes at discrete points in time. The system consists of objects, called *entities*, that have certain properties, called *attributes*. The state of the system is defined as a collection of attributes or state variables that represents the entities of the system. The state may change by the occurrence of an event [9]. In this research, the entities are the Formula One drivers and their race cars. Each driver has certain attributes, such as a DNF probability, a pit stop strategy, and parameters regarding the fuel consumption, the tire degradation and the lap time variability. The state of the system is defined by the order of the drivers on the circuit and the cumulative lap time of each driver, which is equal to the sum of the lap times of all completed laps. The state of the system can only change at the end of each lap. Events that can change the state of the system include differences in lap times between drivers and on-track events.

The simulation model uses the following input data:

- The total number of laps.
- The starting grid.
- The number of positions gained or lost at the start for each individual driver in previous races from the 2016 season.

- The DNF status for each individual driver in previous races from the 2016 season.
- The set of expected pit stop strategies for each individual driver.
- The average pit stop duration of each Formula One team during that race in the 2015 season.
- Driver parameters regarding the fuel consumption, the tire degradation and the lap time variability. These parameters are estimated for each individual driver using the fuel model and the tire model.

Figure 2 shows the design of the simulation model. A simulation starts with simulating which drivers will not finish the race. This is modeled as a realization of the Bernoulli distribution using the DNF probability as probability of success, as will be described in Section 4.3. Next, if a driver retires, the lap that the driver retires in is drawn arbitrarily. The probability of retiring in the first lap is set equal to a relatively high probability, while all other laps are given an equal probability. For each driver, the expected pit strategy is chosen arbitrarily from a set of possible pit strategies. Section 4.6 describes how this set of possible pit strategies is determined. The mixture of cars at the start of the race is modeled in the first lap. The number of positions lost or gained by each driver is determined by drawing a position change from an empirical distribution function. The new position of the drivers is computed by adding the position change to the position on the starting grid. Section 4.4 will describe these empirical distribution functions.

For each lap, the simulation model first determines whether any of the drivers retire in that particular lap. That drivers are removed from the simulation and the safety car is deployed with the safety car probability. The safety car will be present on the circuit during the safety car period. Then a lap time is simulated for each individual driver using the qualifying time, the parameters regarding the fuel consumption and the tire degradation, and the estimated lap time variability. Section 4.2 discusses how the parameters regarding the fuel consumption and tire degradation are estimated. The fuel level can be determined using the lap number and the total number of laps. The expected pit stop strategy is used to determine the tire compound in that particular lap and the age of the tires. In each lap, we also compute a cumulative lap time for each driver, which equals the sum of the lap times of all completed laps. The overtaking model adds interactions between drivers to the simulation model. After individual lap times of drivers are simulated, the overtaking model computes the difference in cumulative lap time between each consecutive pair of drivers on the circuit. The overtaking model determines whether any driver was able to overtake another driver, as will be described in Section 4.5.

Finally, the expected pit stop strategies are used to determine which drivers do a pit stop in the particular lap. Since, the pit lane is often located at the start-finish, pit stops are the last on-track events that are added at the end of each lap. The pit stop time is estimated using pit stop data of the 2015 season for each Formula One team and added to the cumulative lap time. The drivers are sorted by their cumulative lap time at the end of each lap.

We conclude this section by summarizing the model assumptions:

- The Formula One cars do not experience the effects of air resistance. We are not able to model the air resistance as input parameter of the simulation model because of the lack of available data.
- The fuel consumption of each car remains constant for the duration of a race, while the tire degradation is a quadratic function of the number of laps that is driven on the tire compound.
- We make no distinction in different causes of retirements of cars.
- The performance of the driver is reflected in the qualifying time.
- The weather conditions remain dry for the duration of a race.
- Only one car may be overtaken at a time, except for the first lap during the mixture of cars after the start.

- The top 10 drivers choose a pit stop strategy from the set of expected pit strategies, while the other drivers follow the pit stop strategy from last year's winner.

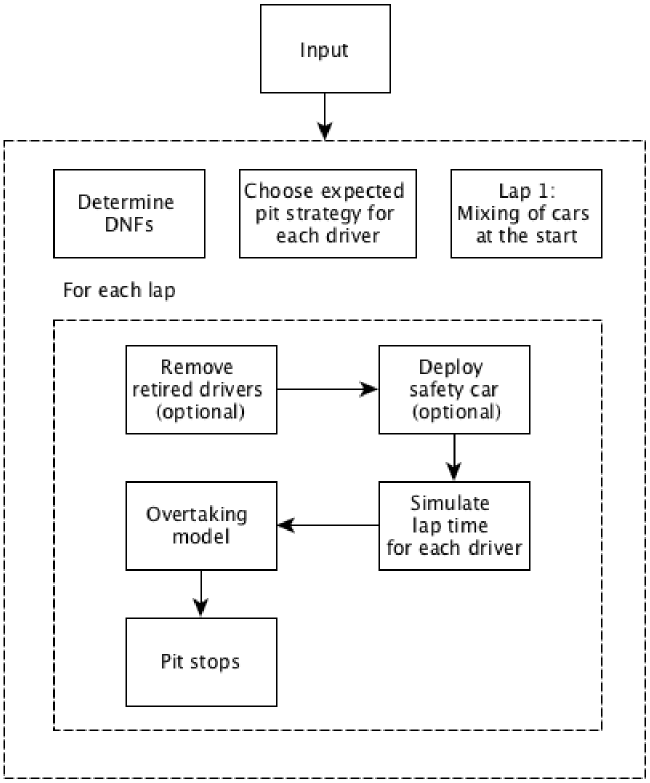

Input
Determine DNFs
Choose expected pit strategy for each driver
Lap 1: Mixing of cars at the start
For each lap
Remove retired drivers (optional)
Deploy safety car (optional)
Simulate lap time for each driver
Overtaking model
Pit stops

Figure 2: Design of the simulation model.

4.2 Modeling lap times

In modeling lap times, we assume that the lap time consists of a base lap time, a fuel bonus, a tire penalty, and a normally distributed random variable representing the lap time variability:

$$\text{(lap time)} = \text{(base lap time)} + \text{(fuel correction)} + \text{(tire correction)} + \text{(random variability)}.$$

The base lap time is equal to the fastest time a driver achieves during the qualifying session. This lap time is achieved with as little fuel as possible on board and on a new warmed-up set of tires. We assume the base lap time to be the minimum lap time a driver can achieve, since the car has higher fuel load, aging tires, and interactions with other cars during the race. The performance of the driver and the car are thus represented in this base lap time. The parameters that describe the influence of fuel consumption and tire degradation on the lap times are estimated for each driver separately based on lap time data of all races in the training set. However, we cannot use the actual lap times to estimate these parameters, because the actual lap times depend on the length of a circuit and its design. Therefore, the actual lap times are reduced by the fastest qualifying time of a driver and these 'corrected lap times' are used instead of the actual lap times. In the following sections, the actual lap times are denoted by Y, while the corrected lap times are denoted by $\tilde{Y}$. Moreover, safety car laps and other outliers are removed from the lap time dataset.

Section 4.2.1 describes the linear regression model that is used to estimate the influence of fuel consumption on the lap times. Then, in Section 4.2.2, a quadratic function is estimated on the residuals of the fuel model to estimate the influence of the tire degradation. Finally, the standard deviation of the random variable representing the lap time variability is estimated by the standard deviation of the residuals of the fuel model and the tire model. We assume this random variable to have a normal distribution with mean equal to 0 and standard deviation equal to the estimated one, as is described in Section 4.2.3.

4.2.1 Fuel model

The fuel model is used to estimate the influence of fuel consumption on the lap times. A linear regression model is estimated for each driver separately. This is because the intercept of the fuel model represents the average difference between the lap times of the race and the qualifying time, which is driver dependent. Also, the influence of the fuel consumption on the lap time depends on the design of the car. We define the following variables for each driver:

$$\begin{aligned}
Y_{it} &= \text{lap time in lap } t \text{ on circuit } i \text{ (in seconds)}\\
Q_i &= \text{fastest qualifying time on circuit } i \text{ (in seconds)}\\
\tilde{Y}_{it} &= Y_{it} - Q_i = \text{corrected lap time in lap } t \text{ on circuit } i \text{ (in seconds)}\\
X_{1it} &= \text{percentage of remaining fuel at the end of lap } t \text{ on circuit } i = \left(1 - \frac{t}{T_i}\right) \cdot 100\%
\end{aligned}$$

where $t = 1, \ldots, T_i$ and T_i is equal to the total number of laps on circuit i. The corrected lap time as a function of the fuel level can be estimated using a linear regression model:

$$\tilde{Y}_{it} = \beta_0 + \beta_1 X_{1it} + e_{it},$$

where β_0 and β_1 are coefficients, and e_{it} is the error in the t^{th} corrected lap time. The coefficient β_1 describes the influence of the fuel consumption on the lap time. The estimated fuel coefficients for each driver can be found in 6.

4.2.2 Tire model

The tire model is used to estimate the influence of tire degradation on the lap times. In contrast to the fuel model, a quadratic regression model is used. This is based on the assumption that, during a stint on a set of tires, the lap times increase slightly at the beginning of the stint because of the tires being relatively cold and at the end of the stint because of tire degradation. A strictly convex function is thus expected to be more suitable to model the influence on the lap times when tire degradation is high. However, on circuits where tire degradation is not very high, a linear regression can also be suitable to model the influence of tire degradation on the lap times. The tire model is estimated for each driver separately, since tire degradation highly depends on the driving style. We only estimate parameters for the dry-weather compounds ultrasoft, supersoft, soft, medium, and hard, since we assume dry weather conditions during the race. Define the following variables for each driver:

$$\begin{aligned}
\hat{e}_{it} &= \tilde{Y}_{it} - \hat{\beta}_0 - \hat{\beta}_1 X_{1it}\\
&= \text{residual lap time in lap } t \text{ on circuit } i\text{, after estimating the fuel model (in seconds)}\\
X_{2it} &= \text{age of the tires in lap } t \text{ on circuit } i
\end{aligned}$$

where $t = 1, ..., T_i$. The residual lap time as a function of the age of the tires can be estimated using a quadratic regression model:

$$\hat{e}_{it} = \sum_{c \in C} (\beta_{0,c} + \beta_{1,c} X_{2it} + \beta_{1,c} X_{2it}^2) \mathbb{1}_{\{\text{lap on compound c}\}} + u_{it}, \quad u_{it} \sim N(0, \sigma^2)$$

where $\beta_{0,c}$, $\beta_{1,c}$, and $\beta_{2,c}$ are coefficients, $c \in C = \{\text{ultrasoft, supersoft, soft, medium, hard}\}$, and u_{it} is the error in the t^{th} residual lap time. We impose the restriction

$$\beta_2 \geq 0$$

to ensure that the lap time is modeled as a convex or linear function of the age of the tires. Note that when β_2 is (approximately) equal to 0, the lap times are modeled as a linear function of the age of the tires rather than a quadratic (strictly convex) function. The estimated tire coefficients for each driver can be found in 6.

4.2.3 Lap time variability

We combine the fuel model and the tire model into a final model that describes the influence of fuel consumption and the age of the tires on the lap time for each individual driver. This model is given by

$$\begin{aligned} Y_{it} &= Q_i + \beta_0 + \beta_1 X_{1it} + \sum_{c \in C} (\beta_{0,c} + \beta_{1,c} X_{2it} + \beta_{1,c} X_{2it}^2) \mathbb{1}_{\{\text{lap } t \text{ on compound } c\}} + u_{it}, \\ u_{it} &\sim N(0, \sigma^2) \end{aligned}$$

where the variables are defined as mentioned before.
We assume the lap time variability to be a normally distributed random variable with mean equal to 0 and standard deviation estimated by the standard deviation of the residuals $\hat{u}_{it}$. The estimated standard deviation $\hat{\sigma}$ for each driver can be found in 6.

Figure 3 compares the estimated lap times based on the final model with Nico Rosberg's actual lap times during the 2016 European Grand Prix. The stint on supersoft tires (laps 1-21) is estimated to be faster than the stint on soft tires (laps 22-51). This is realistic, since it is known that drivers are able to drive faster on softer tires with more grip. However, the lap times during the stint on supersoft tires seem to be structurally underestimated. This is caused by the fact that the parameters of the supersoft tires are estimated using all stints on supersoft tires from the training set. The lap times within other stints were probably lower than the lap times within the depicted stint, causing the underestimation. Nevertheless, the fit of the model is decent.

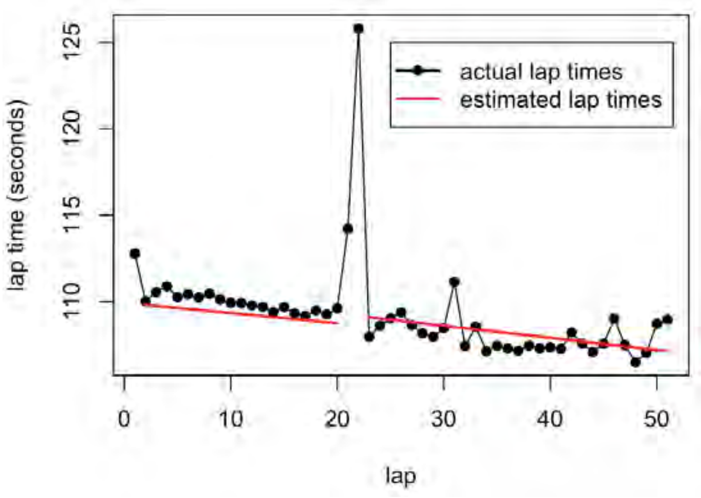


Figure 3: Estimated lap times by the fuel and tire model for Nico Rosberg during the European Grand Prix 2016.

4.3 Crashes, 'Did Not Finish' and safety cars

We model the retirements of drivers during the race by estimating a 'Did Not Finish' (DNF) probability for each driver based on the finishing status in races from the training set. We make no distinction in different causes of retirements of drivers, such as crashes or mechanical failures. One method to estimate the DNF probability of each driver is to use the fraction of not finished races of the total number of races from the training set. However, the DNF probability is then estimated using a small sample. Also, drivers that have finished all the races in the training set are assigned a DNF probability equal to 0, while intuitively there is still a small probability of retiring. Bayesian inference [11] allows us to use the finishing status of all drivers to estimate the finishing status of a particular driver.

Bayesian inference requires a probability model to describe the experimental data and a prior distribution. The probability model here is that of independent Bernoulli trials, since each driver has exactly two outcomes for each race: finishing or not finishing. The number of not finished races for each driver thus follow independent binomial distributions

$$y_j \sim \text{Bin}(n_j, \theta_j),$$

where n_j represents the number of observations for driver j, which is equal to the number of races in the training set. The parameter θ_j represents the DNF probability of driver j.
The aim of Bayesian inference is to update the prior beliefs on the parameter θ_j based on new data. We believe the following on the parameters θ_j:

1. The parameter θ_j represents the DNF probability of driver j. As a result, we believe that $\theta_j \in [0, 1]$ for all j.
2. We expect a right skewed prior distribution for the parameter θ_j, since it is more likely that the driver never retires than that the driver always retires.

These prior beliefs should be quantified in a prior distribution, but finding a prior distribution is in general not straightforward. A non-informative prior, for example, would not reflect our strong beliefs about the shape and the domain of the parameter θ_j. A Beta distribution, however, has a more flexible shape and is defined on the interval $[0, 1]$. The Beta distribution has another important property: the Beta distribution is the conjugate prior of the Bernoulli likelihood. A conjugate prior is a choice of a prior distribution that, when coupled to a specific type of likelihood function, provides a posterior distribution that is of the same family of the prior distribution. When a Beta prior distribution is coupled to a Bernoulli likelihood, then the posterior distribution is given by a Beta distribution with certain parameters. We can thus simply specify the posterior distribution of the parameters θ_j using this property of conjugate priors.

First, we have to estimate the parameters α and β of the prior distribution. These parameters are estimated by the method of moments using the finishing status of all drivers in all races from the training set. The fraction of not finishing is computed for each driver separately. The average of these fractions is used as an estimate of the mean of the prior distribution, while the sample standard deviation is used as an estimate of the standard deviation of the prior distribution. These were computed as $\hat{\mu} = 0.179$ and $\hat{\sigma} = 0.093$, respectively. Since we know that the mean and the standard deviation of the Beta distribution are given by

$$\mu = \frac{\alpha}{\alpha + \beta} \quad \text{and} \quad \sigma = \sqrt{\frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}},$$

we can estimate the value of the parameters α and β by

$$\hat{\alpha} = \left(\frac{1 - \hat{\mu}}{\hat{\sigma}^2} - \frac{1}{\hat{\mu}}\right)\hat{\mu}^2 \quad \text{and} \quad \hat{\beta} = \hat{\alpha}\left(\frac{1}{\hat{\mu}} - 1\right),$$

respectively. This results in the parameter estimates $\hat{\alpha} = 2.856$ and $\hat{\beta} = 13.100$. This method of using the data to estimate the parameters of the prior distribution is called 'empirical Bayes'. Figure 4 shows the probability density function of the Beta(2.856, 13.100) distribution which is used as prior distribution.

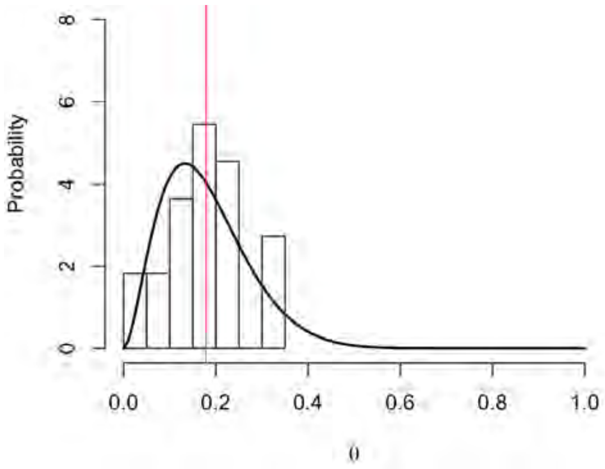


Figure 4: Probability density function of the Beta(2.856, 13.010) distribution plotted against the histogram of the empirical distribution of all DNF probabilities. The mean of the empirical distribution is indicated in red.

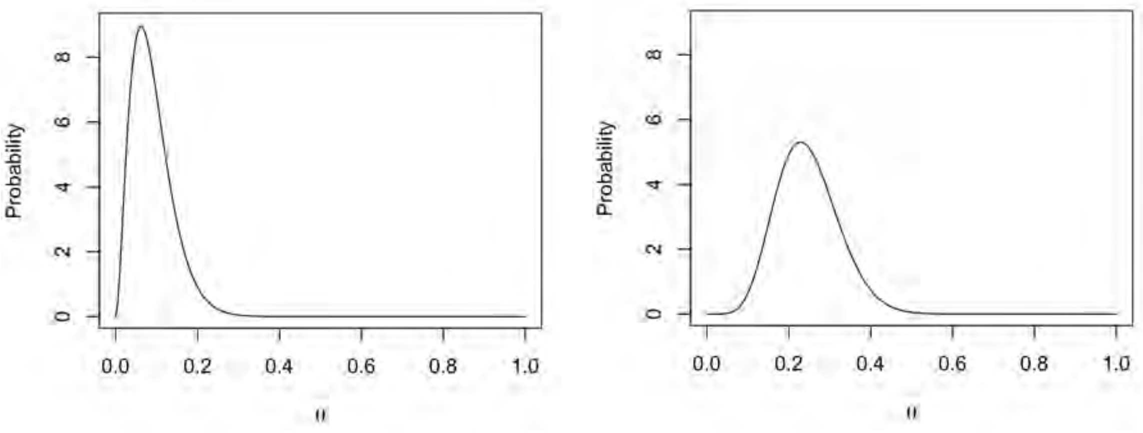
Using Bayes' rule, it can be shown that the posterior distribution of a Beta prior distribution and a Bernoulli likelihood equals a Beta$(z+\alpha, N-z+\beta)$ distribution, where z is the number of successes, N is the number of trials and α and β are the parameters of the prior distribution. We can thus determine the posterior distribution of the DNF probability for each driver using the number of DNFs in the training set. The DNF probability for each driver is estimated by the mean of the posterior distribution. The results are shown in Table 4.

Figure 5 shows the posterior distribution for $z = 0$ and $z = 5$, respectively. This figure illustrates a left-shifted Beta distribution for a relative small number of DNFs and a right-shifted Beta distribution for a relative high number of DNFs, compared to the prior distribution. Bayes inference has thus used the data of all drivers to compute a prior distribution, which is then updated based on driver-specific data resulting in a driver-specific posterior distribution. In particular, a small DNF probability is assigned to drivers that finished all the races in the training set, which is preferred over assigning zero probability.

Next, given the retirement of a driver, we have to determine in which lap the driver retires. We expect the probability that a driver retires in the first lap to be considerably higher than the probability that a driver retires in any other lap, because of the mixture of cars at the start of a race. Therefore, all laps are assigned equal probability of retiring except for the first lap, which is assigned a probability that is ten times larger. When a driver retires, the safety car is deployed with a certain probability during a certain number of laps, called the safety car period. During the safety car period, the simulated lap time of each driver is multiplied by the safety car factor, which is set to 1.2. The safety car is deployed with probability 0.5 during a safety period of 5 laps. In this research, we assume that the safety car is only deployed after the retirement of a driver. The deployment of the safety car because of rainy weather is not included in the model, since we assume dry weather conditions during the race. The deployment of a safety car can highly influence the pit stop strategy of Formula One teams. Formula One Teams often react to a safety car situation by making an early pit stop if the safety car is deployed shortly before a scheduled stop and within the so-called pit window.

Driver	**Number of DNFs (z)**	**Posterior distribution**	**DNF probability (θ_j)**
Nico Rosberg	1	Beta(3.856, 28.100)	0.121
Lewis Hamilton	2	Beta(4.856, 27.100)	0.152
Sebastian Vettel	4	Beta(6.856, 25.100)	0.215
Kimi Raikkonen	2	Beta(4.856, 27.100)	0.152
Daniel Ricciardo	0	Beta(2.856, 29.100)	0.089
Max Verstappen	2	Beta(4.856, 27.100)	0.152
Felipe Massa	2	Beta(4.856, 27.100)	0.152
Valtteri Bottas	1	Beta(3.856, 28.100)	0.121
Nico Hulkenberg	3	Beta(5.856, 26.100)	0.183
Sergio Perez	0	Beta(2.856, 29.100)	0.089
Kevin Magnussen	3	Beta(5.856, 26.100)	0.183
Jolyon Palmer	5	Beta(7.856, 24.100)	0.246
Daniil Kvyat	5	Beta(7.856, 24.100)	0.246
Carlos Sainz	3	Beta(5.856, 26.100)	0.183
Marcus Ericsson	4	Beta(6.856, 25.100)	0.215
Felipe Nasr	4	Beta(6.856, 25.100)	0.215
Fernando Alonso	3	Beta(5.856, 26.100)	0.183
Jenson Button	5	Beta(7.856, 24.100)	0.246
Pascal Wehrlein	4	Beta(6.856, 25.100)	0.215
Esteban Ocon	3	Beta(5.856, 26.100)	0.183
Romain Grosjean	4	Beta(6.856, 25.100)	0.215
Esteban Gutierrez	3	Beta(5.856, 26.100)	0.183

Table 4: Estimated DNF probability per driver using Bayesian inference.



(a) $z = 0$

(b) $z = 5$

Figure 5: Posterior distribution for different values of z.

4.4 Mixing of the cars at the start

The method that is used to model the mixing of cars at the start of the race is based on the method of Bekker and Lotz (2009). Bekker and Lotz use discrete empirical distributions based on historical data to describe the probability that a driver gains or loses a finite number of positions at the start of the race. In this research, the discrete empirical distribution for each driver is based on the positions gained or lost during each race in the training set. This results in a very sparse empirical distribution where positive probability mass is only assigned to a few possible positions, as can be seen from the histogram in Figure 6. As a consequence, the empirical distribution is not a representative distribution to draw from.

The method of smoothed empirical distributions resolves the problem that the empirical distribution assigns positive probability mass only to a finite number of points [10]. Given a sequence of values $y^{(1)}, \dots, y^{(n)}$, $y^{(i)} \in \mathbb{R}$, the smoothed empirical distribution is defined as a mixture distribution with the following probability distribution function

$$f_X(x) = \frac{1}{n}\sum_{i=1}^{n} f_\sigma(x - y^{(i)}),$$

where f_σ is usually a symmetric unimodal probability density function with expectation 0 and variance σ^2. The smoothed empirical distribution is thus a mixture of probability distribution functions $f_\sigma(x - y^{(1)}), \dots, f_\sigma(x - y^{(n)})$ with variance σ^2 and centered at the values $y^{(1)}, \dots, y^{(n)}$. If we choose the normal distribution with expectation 0 and variance σ^2 as f_σ, then we have

$$f_X(x) = \frac{1}{n\sqrt{2\pi\sigma^2}}\sum_{i=1}^{n} e^{\frac{-(x-y^{(i)})^2}{2\sigma^2}}.$$

To smooth the empirical distribution of positions gained for each driver, we choose f_σ to be a normal distribution with parameter σ^2 equal to 1. The smoothing of the empirical distribution is illustrated in Figure 6. This figure compares a histogram describing the empirical distribution of the number of positions lost or gained by Nico Rosberg with the smoothed empirical distribution. The smoothed empirical distribution is used in the simulation model to draw the number of positions a driver loses or gains at the start. This number of positions is then added to the position on the starting grid of each driver. The resulting position cannot be smaller than 1 or bigger than the total number of drivers. If Nico Rosberg, for example, starts second on the starting grid and gains two positions, then he will emerge in the first position after the first lap.

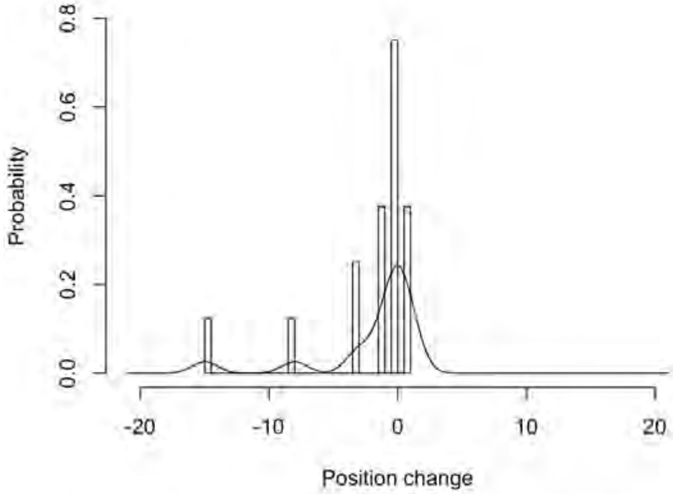


Figure 6: Empirical distribution function describing the positions gained or lost by Nico Rosberg during the start (based on the training set) compared to the smoothed empirical distribution.

4.5 Overtaking model

The model specified so far simulates lap times for each individual driver, but completely ignores interactions between drivers during the race. The overtaking model is used to model these interactions. The most important interactions between drivers are overtaking and crashing. This section describes how overtaking is included in the simulation model, while modeling crashes (retirements) has already been discussed in Section 4.3.

Consider two drivers that drive closely behind each other. If the trailing car is faster than the leading car, then the trailing car can try to overtake the leading car. How successful the overtake is, depends on the relative speed, the distance between the two cars, the position on the circuit where the overtake is planned and the defense of the leading car. The overtaking model determines whether the trailing car is able to overtake the leading car based on the following parameters:

- DRS bonus (t_{DRS}):
  The trailing car receives a DRS bonus (a time reduction) if the time difference with the leading car is smaller than 1.0 second. However, drivers will not be able to access the DRS for the first two laps of the race or during a safety car period.
- Overtaking threshold ($\alpha_{\text{overtaking}}$):
  The trailing car can only overtake the leading car if the time difference is bigger in absolute value than the overtaking threshold.
- Overtaking probability ($p_{\text{overtaking}}$):
  The overtaking action is successful with a certain probability.
- Overtaking penalty ($t_{\text{overtaking}}$):
  When a driver is successfully overtaken, both the driver in the trailing car and the driver in the leading car receive a 'time penalty'.
- Minimum time difference (δ_{min}):
  The minimum time difference represents the closest a driver is allowed to run behind another driver.

Finally, the overtaking model takes the deployment of a safety car into account, since overtaking is not allowed during a safety car period. We will now discuss the design of the overtaking model.

Define the following variables

$$\begin{aligned}
\hat{Y}_{d_j i,t} &= \text{estimated lap time in lap } t \text{ of driver } d_j \text{ on circuit } i\\
\hat{C}_{d_j i,t+1} &= \tilde{C}_{d_j i,t} + \hat{Y}_{d_j i,t+1}\\
\tilde{C}_{d_j i,t+1} &= \hat{C}_{d_j i,t+1} + \text{ time correction}
\end{aligned}$$

for all drivers $j = 1, \dots, |D|$, where D represents the set of all drivers, $t = 1, ..., T_i$, and T_i is equal to the total number of laps on circuit i. $\hat{C}_{d_j i,t+1}$ represents the simulated cumulative lap time in lap t of driver d_j on circuit i, while $\tilde{C}_{d_j i,t}$ represents the corrected simulated cumulative lap time.

After the individual lap time $\hat{Y}_{d_j i,t+1}$ in lap $t+1$ is simulated for all drivers d_j, the overtaking model determines whether any driver was able to overtake another driver in the simulated lap. First, the individual lap time $\hat{Y}_{d_j i,t+1}$ is added to the corrected cumulative lap time of the previous lap t, given by $\tilde{C}_{d_j i,t}$, for each driver d_j. Then the overtaking model computes the difference in cumulative lap time between each consecutive pair of drivers (d_j, d_k) on the circuit. The difference in cumulative lap time between driver d_j and driver d_k is given by

$$\hat{\delta}_{j,k} = \hat{C}_{d_k i,t+1} - \hat{C}_{d_j i,t+1},$$

where for the positions of the drivers at the end of the lap, which are given by $p(d_j)$ and $p(d_k)$, holds that $p(d_j) < p(d_k)$ and $p(d_k) - p(d_j) = 1$. Thus, the difference in cumulative lap time for each pair of driver is computed as the cumulative lap time of the trailing car minus the cumulative lap time of the leading car.

If the trailing car d_k is faster than the leading car d_j, then it holds that $\hat{\delta}_{j,k} < 0$. However, the trailing car can only overtake the leading car if the difference in cumulative lap time is smaller than the overtaking threshold $\alpha_{\text{overtaking}}$. This can be expressed as

$$\hat{\delta}_{j,k} = \hat{C}_{d_k i,t+1} - \hat{C}_{d_j i,t+1} < \alpha_{\text{overtaking}}.$$

Note that $\alpha_{\text{overtaking}} < 0$, because otherwise the trailing car is not faster than the leading car. If the overtaking action was successful, which occurs with the overtaking probability $p_{\text{overtaking}}$, the cars change positions. Both the leading car d_j and the trailing car d_k receive a 'time penalty' because of the overtaking action, $t_{\text{overtaking}} > 0$ (in seconds). If the trailing car is not faster than the leading car but the difference in cumulative lap time is smaller than the DRS threshold, $\alpha_{DRS} > 0$, then the trailing car receives a DRS bonus, $t_{\text{DRS}} < 0$, in the next lap. Finally, the time difference between every pair of cars at the end of each lap is assumed to be no smaller than the minimum time difference, $\delta_{\text{min}} > 0$. After an overtaking action has occurred, we also take the car before the trailing car and the leading car into consideration, when correcting the cumulative lap times for the minimum time difference. The time correction that is used to compute $\tilde{C}_{d_j i,t+1}$ from $\hat{C}_{d_j i,t+1}$ thus consists of the time penalty because of overtaking, a DRS bonus achieved in the previous lap and a correction to ensure that the time difference is no smaller than the minimum time difference. Figure 7 illustrates how the overtaking model works.

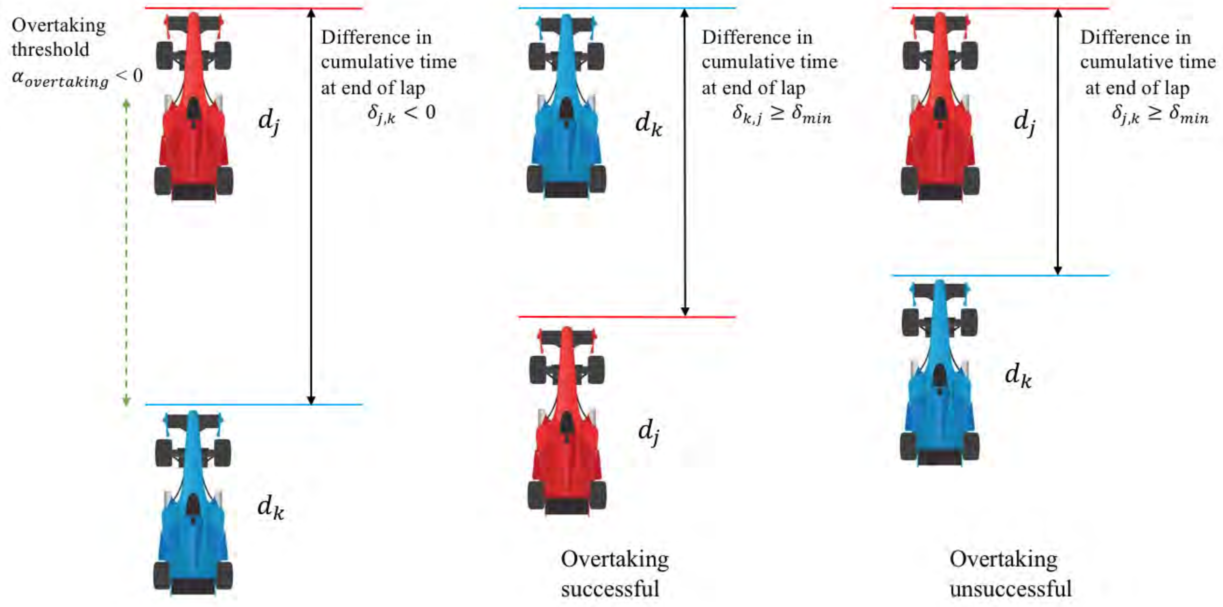

Overtaking threshold
$\alpha_{overtaking} < 0$
Difference in cumulative time at end of lap
$\delta_{j,k} < 0$
d_j
d_k
Difference in cumulative time at end of lap
$\delta_{k,j} \geq \delta_{min}$
d_k
d_j
Overtaking successful
Difference in cumulative time at end of lap
$\delta_{j,k} \geq \delta_{min}$
d_j
d_k
Overtaking unsuccessful

Figure 7: The overtaking model.

The value of the overtaking parameters differ for each of the races in the test set, except for the DRS threshold α_{DRS}, the DRS bonus t_{DRS}, the time penalty after an overtaking action $t_{\text{overtaking}}$ and the minimum time difference δ_{min}. The DRS bonus is chosen in accordance with the regulations of the FIA, which state that a driver within one second of a rival car can activate his DRS within designated DRS activation zones. However, we assume that a driver can activate his DRS in the next lap if the difference in cumulative time is smaller than α_{DRS} at the end of the previous simulated lap, since we have no information whether this difference has exceeded the DRS threshold in a DRS zone or at another segment of the circuit. The value of the DRS bonus t_{DRS}, the time penalty after an overtaking action $t_{\text{overtaking}}$, and the minimum time difference δ_{min} are based on parameters used in the simulation model of *f1metrics* [12]. Table 5 shows the values of these parameters that are used in the simulation model.

Parameter	**Value**
DRS threshold α_{DRS}	1 second
DRS bonus t_{DRS}	-0.5 seconds
Overtaking penalty $t_{\text{overtaking}}$	0.5 seconds
Minimum time difference δ_{min}	0.2 seconds

Table 5: Parameter values used in the overtaking model.

The values of the overtaking threshold $\alpha_{\text{overtaking}}$ and the overtaking probability $p_{\text{overtaking}}$ are chosen by comparing the average number of overtaking actions in the simulation model by the actual number of overtaking actions during the race and presented in Section 5.

4.6 Pit stop strategies

The pit stop strategies are included in the simulation model by selecting a set of expected pit stop strategies for each driver, while taking the tire regulations into account. In each simulation, the pit stop strategy is chosen arbitrarily from the set of expected pit stop strategies for each driver. Also, the lap in which the pit stop is made is determined by drawing from a normal distribution having a mean equal to the planned lap and a standard deviation set to 3. We will now illustrate how the expected pit stop strategies are determined for the Japanese Grand Prix.

According to Pirelli, a two stop strategy seems to be the most probable strategy for the 2016 Japanese Grand Prix. The soft tire, medium tire and hard tire were made available by Pirelli during this race weekend. During qualifying, the soft tire was used from Q2 onwards. As a result, the tire regulations restrict that the top 10 on the starting grid will start the Japanese Grand Prix on soft tires. In addition, the drivers are obliged to have two sets of hard tires available for the race, from which at least one must be used. In 2015, Mercedes driver Lewis Hamilton won with a two-stop strategy, starting on medium tires, changing again to medium tires in lap 16, and then changing to hard tires in lap 31. The best alternative was a medium-hard-hard strategy, used last year by Nico Rosberg leading to a successful undercut, according to Pirelli [13].

Based on this information, we can make some assumptions on the expected tire strategy followed by the drivers during the 2016 Japanese Grand Prix. The expected tire strategy for each driver is contained in Table 6. We assume a two stop pit strategy for each driver. Each driver has to do at least one stint on the hard tire compound and the top 10 drivers have to start the Japanese Grand Prix on the soft tire. The top 10 drivers will thus probably choose one of the following tire strategies: Soft - Hard - Hard, Soft - Hard - Medium, and Soft - Medium - Hard. Notice that the strategy Soft-Medium-Medium is not permitted, since the hard tire is chosen as obligated compound. In each simulation, we arbitrarily choose the tire strategy for each top 10 driver from the set of expected tire strategies. We assume that the other drivers will use Lewis Hamilton's winning strategy during the 2015 Japanese Grand Prix, which was Medium - Medium - Hard. For all drivers, we assume that the first pit stop is planned in lap 16, and the second pit stop is planned in lap 31. The duration of the pit stops is estimated by the average pit stop durations for each team during the 2015 Japanese Grand Prix. The pit stop duration of Renault and Haas F1 Team are chosen equal to the average of all pit stop durations, since they did not yet compete in the Formula One Championship in 2015.

Position	**Driver**	**Tire strategy**
1	Nico Rosberg	Soft - Hard/Medium - Hard/Medium
2	Lewis Hamilton	Soft - Hard/Medium - Hard/Medium
3	Max Verstappen	Soft - Hard/Medium - Hard/Medium
4	Daniel Ricciardo	Soft - Hard/Medium - Hard/Medium
5	Sergio Perez	Soft - Hard/Medium - Hard/Medium
6	Sebastian Vettel	Soft - Hard/Medium - Hard/Medium
7	Romain Grosjean	Soft - Hard/Medium - Hard/Medium
8	Kimi Raikkonen	Soft - Hard/Medium - Hard/Medium
9	Nico Hulkenberg	Soft - Hard/Medium - Hard/Medium
10	Esteban Gutierrez	Soft - Hard/Medium - Hard/Medium
11	Valtteri Bottas	Medium - Medium - Hard
12	Felipe Massa	Medium - Medium - Hard
13	Daniil Kvyat	Medium - Medium - Hard
14	Carlos Sainz	Medium - Medium - Hard
15	Fernando Alonso	Medium - Medium - Hard
16	Jolyon Palmer	Medium - Medium - Hard
17	Kevin Magnussen	Medium - Medium - Hard
18	Marcus Ericsson	Medium - Medium - Hard
19	Felipe Nasr	Medium - Medium - Hard
20	Esteban Ocon	Medium - Medium - Hard
21	Pascal Wehrlein	Medium - Medium - Hard
22	Jenson Button	Medium - Medium - Hard

Table 6: Expected tire strategy for each individual driver.

Formula One team	**Average pit stop duration (in seconds)**
Mercedes AMG Petronas	23.059
Scuderia Ferrari	23.126
Red Bull Racing	23.838
Williams Martini Racing	26.540
Scuderia Toro Rosso	28.704
Sahara Force India F1	24.170
Sauber F1 Team	23.935
McLaren Honda	23.638
Manor Racing	27.103
Renault	24.876
Haas F1 Team	24.876

Table 7: Average pit stop duration for each Formula One team during the 2015 Japanese Grand Prix.

4.7 Model evaluation

The performance of the simulation model is evaluated by simulating the results of the races in the test set. The test set consists of the Japanese Grand Prix, the Mexican Grand Prix, the Grand of the United States and the Grand Prix of Abu Dhabi. The performance of the simulation model is measured using three output parameters of the simulation model:

1. The average number of successful overtaking actions during a race.
2. The average position of each driver at the end of a race.
3. The average race time of each driver at the end of a race.

The average positions and average race times can be used to validate the fuel and tire model, while the average positions and average number of successful overtaking actions can be used to validate the overtaking model. However, some difficulties arise because the output parameters of the simulation model depend on the pit stop strategies, retirements, safety cars and mixing of the cars at the start of the race. Therefore, the simulation model is adjusted to be able to validate the fuel and tire model and the overtaking model in a proper way. The model is adjusted in such a way that not the estimated values but the actual values are used by simulating the races in the test set. The output parameters of the model are evaluated given the actual pit stop strategies, the actual DNF's and the actual safety cars situations. The set of expected pit stop strategies is replaced by the actual pit stop strategy, drivers are removed from the simulated race in the lap they retire and the safety car occurs in actual safety car laps. However, we do not use the actual number of positions gained at the start of the race because we do also not include the actual overtaking actions. We discuss the methods that are used to evaluate the simulation model in more detail in this section. These methods are also used by Bekker and Lotz (2009). Finally, we illustrate how the simulation model can be used by Formula One teams to compare different race strategies.

4.7.1 Evaluation of the simulated overtaking actions

The average number of successful overtaking actions is compared to the actual number of overtaking actions to determine an appropriate overtaking threshold $\alpha_{\text{overtaking}}$ and overtaking probability $p_{\text{overtaking}}$ for each of the races in the test set. These parameters are determined by trial and error. The actual and average number of overtaking actions do not include:

- Position changes on the first lap of the race due to the mixture of cars at the start of the race.
- Position changes due to drivers lapping drivers in the back of the field.
- Positions gained in the pits during the pit stops.
- Positions gained when a car has retired because of a crash or mechanical failure.

4.7.2 Evaluation of the simulated positions

The Spearman rank correlation coefficient is used to compare the simulated positions of the drivers at the end of the race with the actual positions. The simulated position of a driver is equal to the average of the positions of all simulations. The Spearman correlation coefficient is defined as the Pearson correlation coefficient between ranked variables. We test the null hypothesis that there is no correlation between the simulated and actual positions at a significance level of $\alpha = 0.05$:

$$H_0 : \rho = 0$$
$$H_1 : \rho \neq 0,$$

where ρ denotes the Pearson correlation coefficient between the ranked variables. We define the following variables

$$\begin{aligned}
X_i &= \text{actual position of driver } i,\\
Y_i &= \text{simulated position of driver } i,\\
rg(X_i) &= \text{ranking of driver } i \text{ in the vector containing the actual positions},\\
rg(Y_i) &= \text{ranking of driver } i \text{ in the vector containing the simulated positions, and}\\
n &= \text{number of observations.}
\end{aligned}$$

Then, the rank correlation coefficient can be computed as

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)},$$

where $d_i = rg(X_i) - rg(Y_i)$ and n the number of observations. We reject the null hypothesis if and only if the p-value is smaller than 0.05. Rejection of the null hypothesis results in the conclusion that the actual and simulated position are correlated, meaning that the ranking of the car positions is not random.

4.7.3 Evaluation of the simulated race times

The simulated race time of a driver is computed as the average race time of all simulations. We consider pairs (X_i, Y_i) where X_i corresponds to the actual race time of driver i and Y_i corresponds to the simulated race time of driver i. Since the observations are paired, and the respective sets of observations are not expected to be normally distributed, the Wilcoxon matched-pairs test or Wilcoxon signed rank test is applied to test whether the two samples (actual race times and simulated race times) differ significantly. The test assumes that the differences of matched pairs are centered around a common median and that the underlying distribution is symmetric. The following hypothesis is tested at $\alpha = 0.05$:

$$\begin{aligned}
H_0 &: \theta_1 = \theta_2,\\
H_1 &: \theta_1 \neq \theta_2,
\end{aligned}$$

where θ_1 denotes the median of the actual race times and θ_2 denotes the median of the simulated race times. This is equivalent to testing whether the median of the distribution of the differences $Z_i = Y_i - X_i$ is significantly different from 0. So the hypothesis can be rewritten as

$$\begin{aligned}
H_0 &: m_z = 0,\\
H_1 &: m_z \neq 0.
\end{aligned}$$

where m_z denotes the median of the differences of matched pairs. Let $\tilde{Z} = Z_i - m_z$ and let $(R_1, \ldots, R_n)$ denote the vector of ranks of $|\tilde{Z}_1|, \ldots, |\tilde{Z}_n|$ in the corresponding vector of order statistics. This means that $|\tilde{Z}_i|$ is the R_i-th in size (in increasing order) of the $|\tilde{Z}_1|, \ldots, |\tilde{Z}_n|$. Then the following test statistic can be constructed:

$$V = \sum_{i=1}^{n} R_i \text{sgn}(Z_i - m_z)$$

Relatively large values of V indicate that the true distribution of $Z_1, \ldots, Z_n$ has a larger point of symmetry than 0, whereas relatively small values of V indicate a smaller point of symmetry than 0. Note that here we do not want to reject the null hypothesis. Not rejecting the null hypothesis implies that there is not enough statistical evidence that the actual and simulated race times differ.

4.8 Comparing pit stop strategies

Finally, we illustrate how the simulation model can be used by Formula One teams to compare different race strategies. Before a race, the qualifying lap times, the starting grid, and the fuel and tire parameters can be used as input for the simulation model to simulate the race results. The DNF probabilities and the safety car situations are simulated according to the methods described in Section 4.

We conduct a Wilcoxon two-sample test, also known as the Wilcoxon rank sum test or the Mann Whitney test, to test whether the distribution of simulated position for two different pit stop strategies differ significantly. We use this nonparametric test instead of a two sample t-test because we expect the normality assumption of the t-test not to be valid for the distribution of positions. The distribution of positions is more likely to be skewed to the right. Let $X_1, \ldots, X_n$ be the simulated positions using strategy X and $Y_1, \ldots, Y_n$ be the simulated positions using strategy Y, where n is equal to the number of simulations that all drivers in the team has finished. We combine the two samples into one sample $X_1, \ldots, X_n, Y_1, \ldots, Y_n$ of size $N = 2n$ and let $R_1, \ldots, R_n$ be the ranks in the combined sample. The Wilcoxon two-sample test is based on the test statistic

$$W = \sum_{i=1}^{n} R_i.$$

The null hypothesis that the two samples were selected from populations having the same distribution is rejected for large and small values of W. We also create boxplots of the distribution of the simulated positions. We use these methods to find the optimal race strategy for the Mercedes team, as well as for their drivers Nico Rosberg and Lewis Hamilton for the 2016 Japanese Grand Prix.

5 Results

This section presents the results that are generated by the simulation model. First, we give an overview of the model parameters that are used to simulate the races in the test set. Then we evaluate the performance of the simulation model based on the average number of overtakes, the average positions and the average race times. Finally, we illustrate how the simulation model can be used by Formula One teams to compare different race strategies.

5.1 Model parameters

Table 8 contains the model parameters that are used in the simulation model for each race in the test set.

Parameter	**Symbol**	**Japan**	**United States**	**Mexico**	**Abu Dhabi**
Number of laps	T_i	53	56	71	55
DRS bonus (seconds)	t_{DRS}	-0.5	-0.5	-0.5	-0.5
DRS threshold (seconds)	α_{DRS}	1	1	1	1
Overtaking threshold (seconds)	$\alpha_{\text{overtaking}}$	-1.9	-1.8	-1.9	-1.5
Overtaking probability	$p_{\text{overtaking}}$	0.18	0.4	0.1	0.3
Overtaking penalty (seconds)	$t_{\text{overtaking}}$	0.5	0.5	0.5	0.5
Minimum gap length (seconds)	$\delta_{\min}$	0.2	0.2	0.2	0.2
Safety car probability	$p_{\text{safety\_car}}$	0.2	0.2	0.2	0.2
Safety car period (laps)	$n_{\text{safety\_car}}$	5	5	5	5
Multiplying factor safety car	$C_{\text{safety\_car}}$	1.2	1.2	1.2	1.2

Table 8: Parameter values used in the simulation model.

5.2 Evaluation of the simulated overtaking actions

Table 9 compares the average number of overtaking actions over 1000 simulations with the actual number of overtaking actions per race. This table shows that the Mexican Grand Prix had the lowest number of overtaking actions (24.48) while the Grand Prix of the United States had the highest number of overtaking actions (56.86). This explains the difference in the value of the overtaking parameters. We use a relatively high (in absolute value) overtaking threshold and a relatively low overtaking probability while simulating the Mexican Grand Prix. This results in a relatively low average number of overtaking actions, close to the actual number. On the other hand, we use a relatively low (in absolute value) overtaking threshold and a relatively high overtaking probability while simulating the Grand Prix of the United States, resulting in a relatively high number of overtaking actions. Note that using the actual number of overtaking actions can lead to overfitting, but unfortunately there is no data available about overtaking actions during the 2015 season.

	Actual number of overtakes	**Average number of overtakes**
Japan	43	44.68
United States	55	56.86
Mexico	25	24.48
Abu Dhabi	41	42.64

Table 9: Average number of overtakes compared to the actual number of overtakes per race.

5.3 Evaluating the simulated positions and race times

This section evaluates the performance of the simulation model using the actual pit stop strategies, actual retirements and actual safety car situations. Tables 10, 11, 12 and 13 compare the actual positions with the simulated positions and the actual race times with the simulated race times for each race in the test set. These results are obtained by running 1000 simulations. Each table also contains the results of the Spearman rank correlation test and the Wilcoxon signed rank test. The p-value of the Spearman rank correlation test is smaller than 0.05 for each of the races in the test set, which means that the actual and estimated positions are highly correlated. The highest correlation is achieved while simulating the Grand Prix of Abu Dhabi, while the lowest one is achieved while simulating the Grand Prix of the United States. The results of the Wilcoxon matched-pairs test differ for the races in the test set. Only for the Japanese Grand Prix we cannot reject the null hypothesis that the actual race time and simulated race time have the same underlying distribution. The actual race times of the Grand Prix of the United States and the Grand Prix of Abu Dhabi are underestimated, while those of the Mexican Grand Prix are overestimated.

		Position		**Race time**	
Driver	**Laps**	**Actual**	**Simulated**	**Actual**	**Simulated**
Lewis Hamilton	53	3	1.68	5209.11	5189.47
Nico Rosberg	53	1	3.11	5203.33	5229.80
Max Verstappen	53	2	4.70	5208.31	5253.35
Sebastian Vettel	53	4	4.75	5223.60	5255.83
Daniel Ricciardo	53	6	6.43	5237.27	5278.78
Valtteri Bottas	53	10	6.48	5301.66	5284.04
Felipe Massa	53	9	6.77	5301.10	5296.06
Kimi Raikkonen	53	5	7.49	5231.70	5290.03
Romain Grosjean	53	11	7.82	5302.59	5293.79
Sergio Perez	53	7	7.97	5260.83	5299.94
Nico Hulkenberg	53	8	8.80	5262.51	5304.90
Jolyon Palmer	52	12	13.09	5219.81	5207.80
Fernando Alonso	52	16	14.70	5238.42	5213.09
Kevin Magnussen	52	14	14.99	5235.52	5215.05
Marcus Ericsson	52	15	15.56	5236.28	5223.73
Felipe Nasr	52	19	15.57	5251.58	5220.58
Esteban Gutierrez	52	20	15.73	5254.65	5218.16
Daniil Kvyat	52	13	16.96	5225.85	5225.17
Carlos Sainz Jr.	52	17	18.68	5239.31	5228.99
Jenson Button	52	18	19.43	5242.78	5231.44
Pascal Wehrlein	52	22	20.88	5276.64	5243.27
Esteban Ocon	52	21	21.42	5256.78	5251.09
Test statistic		$r_s = 0.932$		$V = 125$	
P-value		4.119×10^{-6}		0.975	
Decision		Reject H_0		Do not reject H_0	

Table 10: Actual results of the Japanese Grand Prix compared to the simulated results.

		Position		Race time	
Driver	**Laps**	**Actual**	**Simulated**	**Actual**	**Simulated**
Lewis Hamilton	56	1	1.44	5892.62	5793.32
Nico Rosberg	56	2	2.28	5897.14	5810.71
Daniel Ricciardo	56	3	2.56	5912.31	5815.57
Sebastian Vettel	56	4	3.91	5935.75	5860.64
Carlos Sainz Jr.	56	6	5.19	5988.74	5905.16
Fernando Alonso	56	5	5.61	5986.57	5919.84
Daniil Kvyat	55	11	8.37	5947.44	5811.37
Marcus Ericsson	55	14	9.10	5958.59	5827.66
Felipe Massa	55	7	9.44	5899.09	5823.97
Felipe Nasr	55	15	10.70	5975.36	5831.89
Valtteri Bottas	55	16	11.72	5985.67	5840.26
Sergio Perez	55	8	11.99	5913.66	5839.71
Jolyon Palmer	55	13	12.13	5953.77	5839.94
Kevin Magnussen	55	12	12.59	5948.13	5840.37
Romain Grosjean	55	10	13.01	5923.52	5842.99
Jenson Button	55	9	16.13	5915.62	5875.55
Pascal Wehrlein	55	17	16.83	5986.82	5893.18
Esteban Ocon	54	18	18.00	5916.49	5791.06
Kimi Raikkonen	38	19	19.00	4052.75	4012.95
Max Verstappen	28	20	20.00	2968.12	2941.21
Esteban Gutierrez	16	21	21.00	1768.94	1720.44
Nico Hulkenberg	1	22	22.00	143.18	105.91
Test statistic		$r_s = 0.886$		$V = 253$	
P-value		2.853×10^{-6}		4.768×10^{-7}	
Decision		Reject H_0		Reject H_0	

Table 11: Actual results of the Grand Prix of the United States compared to the simulated results.

		Position		**Race time**	
Driver	**Laps**	**Actual**	**Simulated**	**Actual**	**Simulated**
Nico Rosberg	71	2	2.52	6039.76	6180.97
Max Verstappen	71	4	3.24	6052.73	6186.98
Lewis Hamilton	71	1	3.73	6031.40	6214.65
Sebastian Vettel	71	5	3.75	6048.72	6200.13
Nico Hulkenberg	71	7	3.84	6090.29	6200.70
Kimi Raikkonen	71	6	5.46	6080.78	6225.27
Valtteri Bottas	71	8	7.34	6097.01	6274.15
Felipe Massa	71	9	7.85	6107.61	6298.05
Sergio Perez	71	10	8.40	6108.20	6300.82
Daniel Ricciardo	71	3	8.87	6052.26	6295.39
Carlos Sainz Jr.	70	16	12.21	6097.26	6212.18
Fernando Alonso	70	13	12.41	6081.40	6198.80
Jenson Button	70	12	12.95	6072.96	6230.78
Jolyon Palmer	70	14	14.49	6087.60	6252.46
Marcus Ericsson	70	11	15.17	6066.18	6255.43
Kevin Magnussen	70	17	15.54	6101.27	6244.53
Felipe Nasr	70	15	15.73	6095.46	6247.98
Esteban Gutierrez	70	19	18.66	6101.93	6293.26
Romain Grosjean	70	20	18.79	6110.94	6290.06
Daniil Kvyat	70	18	19.06	6101.27	6308.16
Esteban Ocon	69	21	21.00	6061.12	6295.17
Pascal Wehrlein	0	22	22.00	-	0.00
Test statistic		$r_s = 0.9322$		$V = 0$	
P-value		4.119×10^{-6}		9.537×10^{-7}	
Decision		Reject H_0		Reject H_0	

Table 12: Actual results of the Mexican Grand Prix compared to the simulated results.

		Position		**Race time**	
Driver	**Laps**	**Actual**	**Simulated**	**Actual**	**Simulated**
Max Verstappen	55	4	2.21	5885.70	5833.94
Lewis Hamilton	55	1	2.30	5884.01	5837.27
Nico Rosberg	55	2	3.60	5884.45	5861.50
Daniel Ricciardo	55	5	4.26	5889.33	5868.25
Kimi Raikkonen	55	6	4.50	5902.83	5873.52
Sebastian Vettel	55	3	5.56	5884.86	5895.49
Sergio Perez	55	8	7.54	5942.79	5928.24
Nico Hulkenberg	55	7	7.68	5934.13	5929.07
Romain Grosjean	55	11	8.93	5960.79	5949.14
Fernando Alonso	55	10	9.54	5943.91	5955.89
Felipe Massa	55	9	9.96	5943.45	5960.28
Esteban Gutierrez	55	12	11.93	5979.13	5997.99
Marcus Ericsson	54	15	13.68	5921.67	5877.64
Jolyon Palmer	54	17	13.77	5935.20	5873.67
Pascal Wehrlein	54	14	15.18	5920.78	5888.70
Felipe Nasr	54	16	15.74	5927.04	5895.85
Esteban Ocon	54	13	16.63	5914.47	5907.84
Carlos Sainz Jr.	41	18	18.00	4522.13	4512.53
Daniil Kvyat	14	19	19.00	1562.92	1560.30
Jenson Button	12	20	20.00	1362.84	1305.70
Valtteri Bottas	6	21	21.00	663.07	657.15
Kevin Magnussen	5	22	22.00	587.31	575.41
Test statistic		$r_s = 0.965$		$V = 215$	
P-value		3.598×10^{-6}		0.003	
Decision		Reject H_0		Reject H_0	

Table 13: Actual results of the Grand Prix of Abu Dhabi compared to the simulated results.

5.4 Comparing pit stop strategies

This section illustrates how the simulation model can be used by Formula One teams to determine the optimal race strategy among a set of possible strategies. We use the simulation model to find the optimal pit stop strategies for the Mercedes teams as well as for the drivers Nico Rosberg and Lewis Hamilton prior to the 2016 Japanese Grand Prix. Table 14 describes the pit stop strategies that we consider. These pit stop strategies are chosen as follows. Strategy 1 is based on Max Verstappen's actual race strategy during the 2016 Japanese Grand Prix. Max Verstappen made an early pit stop compared to the other drivers but used the same tire compounds as most drivers. Strategy 2 is based on Sebastian Vettel's actual race strategy, which ended with a stint on soft tires. Finally, strategy 3 ends with a stint on medium tires according to the actual race strategy of the Force India team.

	Actual strategy	**Simulated strategies**		
		Strategy 1	**Strategy 2**	**Strategy 3**
Nico Rosberg				
Number of planned pit stops	2	2	2	2
Lap numbers of planned pit stops	12, 29	10, 28	12, 29	12, 29
Tyre choice	S-H-H	S-H-H	S-H-S	S-H-M
Lewis Hamilton				
Number of planned pit stops	2	2	2	2
Lap numbers of planned pit stops	13, 33	10, 28	12, 29	12, 29
Tyre choice	S-H-H	S-H-H	S-H-S	S-H-M

Table 14: Simulated race strategies for the Japanese Grand Prix

We performed 1000 simulations for each combination of pit stop strategies. The average positions given the driver has finished (x, y) are shown in Table 15, where x represents the average position of Lewis Hamilton and y represents the average position of Nico Rosberg. Table 16 shows the average of the sum of the positions given both drivers have finished.

		Nico Rosberg			
		Actual	**Strategy 1**	**Strategy 2**	**Strategy 3**
Lewis Hamilton	**Actual**	(1.68, 3.12)	(1.59, 3.66)	(1.76, 2.70)	(1.75, 3.08)
	Strategy 1	(2.47, 2.92)	(2.30, 3.63)	(2.40, 2.62)	(2.36, 3.11)
	Strategy 2	(2.25, 2.99)	(2.21, 3.53)	(2.28, 2.69)	(2.14, 2.91)
	Strategy 3	(2.89, 2.83)	(2.62, 3.61)	(2.82, 2.60)	(2.81, 2.86)

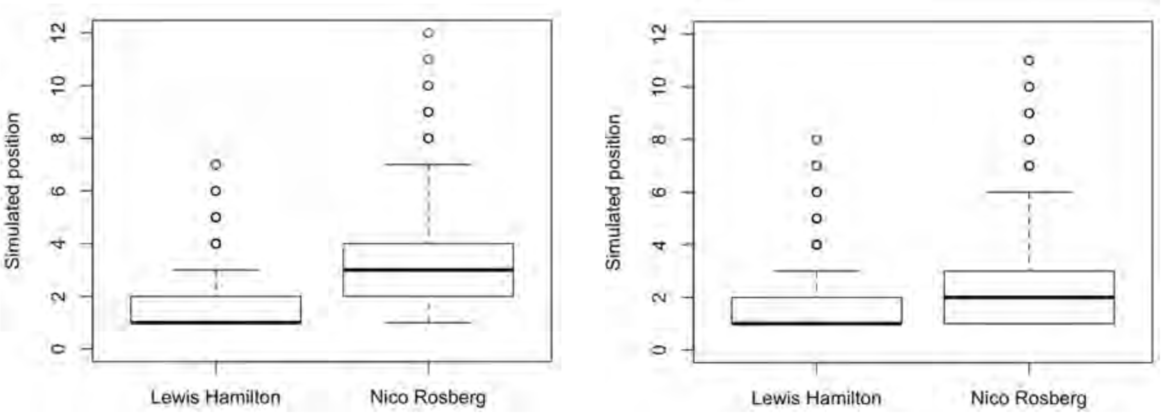
Table 15: Results simulated race strategies for each driver.

		Nico Rosberg			
		Actual	**Strategy 1**	**Strategy 2**	**Strategy 3**
Lewis Hamilton	**Actual**	5.02	5.37	4.59	4.91
	Strategy 1	5.54	6.02	5.21	5.62
	Strategy 2	5.34	5.91	5.09	5.18
	Strategy 3	5.87	6.34	5.55	5.82

Table 16: Results simulated race strategies for the Mercedes team.

From Table 15 we can determine the optimal strategy for each driver, while from Table 16 we can determine the optimal strategy for the Mercedes team. Table 15 suggests that the actual strategy is the optimal strategy for Lewis Hamilton, but that Nico Rosberg should choose strategy 2 or strategy

3 instead of his actual strategy. This can also be seen in the boxplots of Figures 8a and 8b, where we compare the strategy 2 with Nico Rosberg's actual strategy. These boxplots show that Nico Rosberg's median position is smaller for strategy 2 than for the actual strategy. We use the Wilcoxon rank rsum test (Mann Whitney test) to investigate whether the strategy 2 or strategy 3 are indeed better than the actual strategy. Table 17 contains the results. These results show that strategy 2 is significantly better than the other strategies at a significance level of $\alpha = 0.05$, since the null hypothesis is rejected. We can also conclude that strategy 3 is significantly better than the actual strategy. Nico Rosberg should thus choose strategy 2 according to the results of the simulation model before the start of the race. The results of the simulation model do not show any evidence that Lewis Hamilton should choose another strategy than his actual one.



(a) Actual/actual.

(b) Actual/2.

Figure 8: Boxplot of the average positions of the best strategies.

Strategy		Strategy			
Hamilton	**Rosberg**	**Hamilton**	**Rosberg**	**Alternative**	**p-value**
actual	2	actual	actual	less	1.003×10^{-9}*
actual	2	actual	3	less	2.445×10^{-5}*
actual	3	actual	actual	less	0.023*

Table 17: Results of the Wilcoxon rank sum test. Testing for a difference in distribution of Nico Rosberg's simulated positions for multiple race strategies. * denotes rejection at 0.05 significance level.

However, we are interested in the optimal strategy for the Mercedes team rather than the optimal strategy for each driver separately. The optimal strategy for the Mercedes team is given by the strategy that maximizes the positions of both drivers, i.e., the sum of the positions. The results in Table 16 show that the average of the sum of the positions of both drivers in the team is maximized if Nico Rosberg chooses strategy 2 and Lewis Hamilton chooses his actual strategy. As second best strategy Nico Rosberg chooses strategy 3 and Lewis Hamilton chooses his actual strategy. The actual race strategy is the third best strategy. In the following, we will use the notation x/y, where x denotes the strategy of Lewis Hamilton and y denotes the strategy of Nico Rosberg. We now test whether the differences in the average of the sum of the positions are significant. The Wilcoxon rank sum test can again be used to test whether the distribution of the sum of the simulated positions differ significantly for the different strategies. We test the null hypothesis that both strategies are equally good against the alternative hypothesis that one strategy performs better than the other strategy strategy. The results can be found in Table 18. From this table we can conclude that strategy actual/2 performs better than strategy actual/3 since the true location shift of the distribution of the sum of the positions is significantly smaller than 0 at a significance level of $\alpha = 0.05$. Strategy actual/2 is also significantly better than the actual race strategy of the Mercedes team. However, strategy actual/3 is not significantly better than the actual race strategy because the null hypothesis cannot be rejected at a significance level $\alpha = 0.05$.

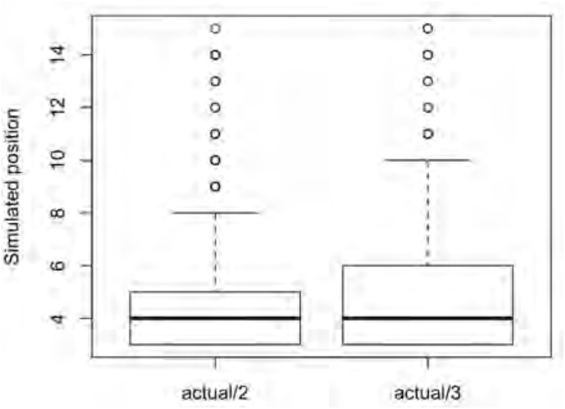


Figure 9: Comparing the average of the sum of the positions of the best strategies.

Strategy		Strategy			
Hamilton	**Rosberg**	**Hamilton**	**Rosberg**	**Alternative**	**p-value**
actual	2	actual	actual	less	1.558×10^{-6}*
actual	2	actual	3	less	8.278×10^{-4}*
actual	3	actual	actual	less	0.065

Table 18: Results of the Wilcoxon rank sum test. Testing for a difference in distribution of the sum of the simulated positions for multiple race strategies. * denotes rejection at 0.05 significance level.

The optimal strategy for the Mercedes team is thus given by strategy 2 for Nico Rosberg in combination with the actual pit stop strategy for Lewis Hamilton according to the results of the simulation model.

6 Discussion and conclusion

The research goal was to build a simulation model that can be used by Formula One teams to determine the optimal race strategy among a set of possible strategies. The optimal race strategy was defined as the strategy that maximizes the positions of both drivers in the team. The two main components of the simulation model, the tire and fuel model and the overtaking model, were validated using races from the test set. To simulate race strategies prior to the race, we also added most on-track events to the simulation model, including the mixing of the cars at the start of a race, pit stops, overtaking actions, safety car situations and driver retirements. Finally, we illustrated how the simulation model can be used by Formula One teams to compare different race strategies.

The performance of the simulation model was evaluated using races from the test set. We compared the average number of overtaking actions, the average position of each driver at the end of the race and the average race time of each driver with the actual values. We can conclude that the simulation model performs well in predicting the actual positions of the drivers but not so well in predicting the actual race times. The actual number of overtaking actions were mainly used to obtain suitable parameter values for the overtaking model.
The high Spearman rank correlation coefficients can imply that the simulated relative difference in performance between drivers, i.e., the relative difference in lap times, and the simulated position changes are close to reality. However, the under- and overestimation of the race times show that the tire and fuel model can be improved. One reason that the simulated race times differ significantly from the actual end times could be that the circuit characteristics are not included in the tire and fuel model. It is known that the degradation of a tire compound and the fuel consumption can differ per circuit. The race times of the Mexican Grand Prix were overestimated by the simulation model. This could be caused by the low tire degradation on the Autódromo Hermanos Rodríguez circuit because of the altitude of the circuit, which is over 2,200 meters above sea level. As a result, the loss in lap time because of tire degradation is probably lower than described by the parameters of the tire and fuel model. Additional research is needed to investigate why the race times of the Grand Prix of the United States and the Grand Prix of Abu Dhabi are underestimated. It could be that the time loss caused by tire degradation is higher than is described by the tire parameters. Another reason could be that drivers lose time because of traffic. It is for example known that overtaking is quite difficult on the Yas Marina circuit where the Grand Prix of Abu Dhabi is organized. As a result, drivers can loose a considerable amount of time by being stuck behind another driver.

By comparing the results of this research to the results of Bekker and Lotz (2009), we can conclude that the simulation model performs almost as well in predicting the end positions but worse in predicting the race times. The Spearman rank correlation coefficients are somewhat lower than those obtained by Bekker and Lotz. However, it is difficult to compare the results because Bekker and Lotz's results are based on 50 simulations instead of 1000. The simulation model of Bekker and Lotz generates more desirable results for the race times than the simulation model in this research. Their results show that there is not enough statistical evidence to reject the null hypothesis that the actual race times and the simulated race times originate from a different distribution for all simulated races. Also, their simulation model does not structurally under- or overestimate the race times, as is the case for the simulation model in this research. This might be caused by the fact that Bekker and Lotz use more circuit specific data to simulate the lap times.

We also illustrated how the simulation model can be used by Formula One teams to determine the optimal race strategy among a set of possible strategies. These results showed that, according to the simulation model, the optimal team strategy prior to the race was given by strategy 2 for Nico Rosberg in combination with the actual pit stop strategy for Lewis Hamilton for the 2016 Japanese Grand Prix. Further research can be done to compare the strategies from a game theoretical point of view, where one could try to find the so-called Nash equilibrium.

Further research can also improve the different components of the simulation model. The fuel and tire model can be improved by including circuit characteristics because the fuel consumption and the tire degradation can differ per circuit. Also, one could use other data sources, such as data from free practice sessions, to estimate the fuel and tire parameters. These data could improve the fuel and tire model because race data might not as clean as data from, for example, free practice sessions. Ideally, you want to use lap time data of drivers in clean air, i.e. not subject to interactions with other drivers, to estimate the tire and fuel model. However, these data is not as easy available as the lap time data from races. The DNF probabilities and the mixing of the cars at the start of the race could be estimated more accurately by adding data from previous seasons. To improve the DNF probabilities one could, for example, give a higher weight to recent retirements then to retirement further in the past. Comparing the frequentist approach with the Bayesian approach can also be useful for further research. To improve the mixing of the cars at the start of the race other methods than the smoothed empirical distribution function can be considered that are more suitable for discrete distribution functions. Finally, more research on the overtaking parameters can improve the overtaking model. One could, for example, try to optimize the overtaking parameters by using grid search or a meta-heuristic.

We can conclude that the simulation model can be used by Formula One racing teams to compare different strategies prior to the race. Simulating race results given a particular strategy can help Formula One teams planning and evaluating their race strategies. However, ideally the simulation model should be able to analyze race strategies in real time such that Formula One teams can adjust their strategies if race incidents occur. Formula One teams indeed use such real time simulation models trying to gain competitive advantages.

References

[1] Higginbotham, S., (November 12, 2015) How Formula 1 Teams Use Big Data to Win [Online] (no place), Fortune.
Available: http://fortune.com/2015/11/12/big-data-formula-1-championship-race/
(Accessed August 4, 2017)

[2] Bekker, J., & Lotz, W. (2009). Planning Formula One race strategies using discrete-event simulation. Journal of the Operational Research Society, 60(7), 952-961.

[3] Formula One overtaking data
Available: http://cliptheapex.com/overtaking/ (Accessed August 21, 2017)

[4] Formula One Database API
Available https://ergast.com (Accessed July 17, 2017)

[5] Pit stop strategy data
For example, pit stop data from the European Grand Prix (2016).
Available: http://www.f1fanatic.co.uk/2016/06/19/2016-european-grand-prix-tire-strategies-and-pit-stops/(Accessed July 17, 2017)

[6] Qualifying session data
For example, qualifying times during the qualifying session of the European Grand Prix (2016).
Available: http://www.fia.com/events/fia-formula-one-world-championship/season-2016/qualifying-classification-6 (Accessed July 17, 2017)

[7] Starting grid data
For example, the starting grid for the European Grand Prix (2016).
Available: https://www.formula1.com/en/results.html/2016/races/958/europe/starting-grid.html (Accessed August 22, 2017)

[8] Race classification data
For example, race classification data of the European Grand Prix (2016).
Available: http://www.fia.com/events/fia-formula-one-world-championship/season-2016/race-classification-6 (Accessed August 22, 2017)

[9] Law, A. M. & Kelton, W. D. (2007). Simulation modeling and analysis (Vol. 3). New York: McGraw-Hill.

[10] Lebanon, G. (2012). Probability: The Analysis of Data, Volume 1, Chapter 3.15.

[11] Gelman, A., et al. Bayesian data analysis. Vol. 2. Boca Raton, FL: CRC press, 2014, pages 29-39.

[12] F1metrics (October 3, 2014) Mathematical and statistical insights into Formula 1, Building a Race Simulator [Online] (no place)
Available: https://f1metrics.wordpress.com/2014/10/03/building-a-race-simulator/ (Accessed May 11, 2017)

[13] Pirelli Grand Prix preview
For example, Pirelli's preview for the Japanese Grand Prix (2016)
Available: http://news.pirelli.com/global/en-ww/japan-2016-preview (Accessed July 31, 2017)

[14] Pirelli Press Release
For example, Pirelli's Press Release about the 2016 Japanese Grand Prix, Qualifying
Available: http://www.pirelli.com/corporate/en/press/2016/10/08/2016-japanese-grand-prix-qualifying/ (Accessed July 31, 2017)

Appendix I

Table 19 shows the estimated parameters for each individual driver.

Driver	**Fuel**	**Ultrasoft**	**Supersoft**	**Soft**	**Medium**	**Hard**	$\hat{\sigma}$
Nico Rosberg	5.595	0.553	0.035	0.076	0.249	0.035	1.480
	0.017	0.002	-0.026	-0.037	0.017	0.034	
		0.000	0.000	0.000	0.000	0.000	
Lewis Hamilton	5.266	-0.085	0.141	-0.457	-0.001	0.493	1.496
	0.022	-0.038	0.068	-0.013	0.072	-0.067	
		0.000	0.000	0.000	0.000	0.000	
Sebastian Vettel	5.180	-1.190	-0.508	0.130	0.704	0.000	1.394
	0.018	0.126	-0.008	-0.012	0.039	0.000	
		0.000	0.000	0.000	0.000	0.000	
Kimi Raikkonen	4.890	-0.524	-0.101	0.326	0.765	0.077	1.243
	0.022	0.087	-0.031	-0.049	0.040	0.055	
		0.000	0.000	0.000	0.000	0.000	
Daniel Ricciardo	4.921	-2.036	-0.292	0.234	0.581	-0.018	1.239
	0.026	-0.006	0.029	-0.037	0.004	0.034	
		0.000	0.000	0.000	0.000	0.000	
Max Verstappen	5.080	-1.962	0.067	0.276	0.563	-0.126	1.432
	0.020	0.020	0.032	-0.053	0.006	0.011	
		0.000	0.000	0.000	0.000	0.000	
Felipe Massa	5.606	1.044	0.059	-0.632	0.398	1.146	1.373
	0.018	-0.044	-0.054	-0.024	0.032	0.015	
		0.000	0.000	0.000	0.000	0.000	
Valtteri Bottas	5.493	-0.977	-0.405	-0.086	0.761	0.078	1.446
	0.022	-0.107	0.029	-0.025	0.030	-0.056	
		0.000	0.000	0.000	0.000	0.000	
Nico Hulkenberg	4.854	-1.898	-0.621	-0.142	1.006	0.729	1.085
	0.026	0.045	-0.031	-0.013	0.005	-0.099	
		0.000	0.000	0.000	0.000	0.000	
Sergio Perez	5.313	-1.334	-0.850	0.046	0.681	-0.212	1.560
	0.022	-0.273	-0.034	-0.025	0.021	-0.014	
		0.000	0.000	0.000	0.000	0.000	
Kevin Magnussen	4.479	-0.573	-0.982	0.137	1.302	0.826	1.230
	0.025	-0.020	0.084	-0.039	-0.020	0.091	
		0.000	0.000	0.000	0.000	0.000	
Jolyon Palmer	4.685	-0.615	-0.690	-0.299	0.540	0.276	1.429
	0.025	-0.146	0.015	-0.007	0.023	0.041	
		0.000	0.000	0.000	0.000	0.000	
Daniil Kvyat	5.167	-0.147	-0.700	-0.496	0.840	0.316	1.365
	0.025	-0.081	0.041	0.014	-0.029	0.009	
		0.000	0.000	0.000	0.000	0.000	
Carlos Sainz	5.070	0.939	-0.237	-0.056	0.982	-0.284	1.491
	0.024	-0.203	0.021	-0.040	0.026	0.045	
		0.000	0.000	0.000	0.000	0.000	
Marcus Ericsson	4.552	-0.628	-0.324	-0.090	0.322	-0.903	1.435
	0.024	0.154	-0.026	-0.013	0.039	0.043	
		0.000	0.000	0.000	0.000	0.000	

Felipe Nasr	4.531	-2.757	-0.174	-0.280	0.697	0.437	1.474
	0.023	0.328	-0.010	-0.026	0.020	0.004	
		0.000	0.000	0.000	0.000	0.000	
Fernando Alonso	4.765	0.934	0.325	-0.445	0.565	-0.823	1.488
	0.024	-0.090	0.005	-0.001	0.012	-0.102	
		0.000	0.000	0.000	0.000	0.000	
Jenson Button	4.598	-1.666	-0.247	-0.247	0.323	0.899	1.285
	0.030	0.116	0.028	-0.040	0.043	-0.002	
		0.000	0.000	0.000	0.000	0.000	1.285
Pascal Wehrlein	4.551	0.163	-1.256	-0.327	-0.246	0.218	1.346
	0.020	0.002	0.121	-0.004	0.103	0.030	
		0.000	0.000	0.000	0.000	0.000	
Esteban Ocon	4.757	1.582	-0.694	-1.140	-0.233	0.126	1.627
	0.023	-0.139	-0.015	0.034	0.122	0.053	
		0.000	0.000	0.000	0.000	0.000	
Romain Grosjean	4.249	-1.863	-0.317	0.464	0.167	0.000	1.316
	0.029	0.101	0.022	-0.031	-0.009	0.000	
		0.000	0.000	0.000	0.000	0.000	1.316
Esteban Gutierrez	4.777	0.617	-0.959	-0.852	1.015	1.732	1.474
	0.028	-0.037	0.018	0.049	0.002	-0.038	
		0.000	0.000	0.000	0.000	0.000	

Table 19: Estimated parameters of the fuel and tire model per driver (rounded to 3 decimals).

Appendix II

Table 20 shows the manually collected data on safety car laps during each race from the 2016 season.

Race	**(Virtual) Safety car laps**
Australia	17 18 19
Bahrein	-
China	4 5 6 7 8
Russia	1 2 3
Spain	1 2 3
Monaco	35 36 37 50 51 68 69
Canada	11
Europe	-
Austria	27 28 29 30 31
Great Britain	1 2 3 4 5 7 8
Germany	-
Belgium	2 3 6 7 8 9 10
Italy	-
Singapore	1 2
Malaysia	1 2 9 10 41 42 43
Japan	-
United States	31 32
Mexico	1 2
Abu Dhabi	-

Table 20: Safety car laps per race.