UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA INDUSTRIAL

OPTIMIZATION OF PIT STOP STRATEGIES IN FORMULA 1 RACING: A DATA-DRIVEN APPROACH

TESIS PARA OPTAR AL GRADO DE MAGÍSTER EN GESTIÓN DE OPERACIONES

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERA CIVIL INDUSTRIAL

MARIANA PAZ QUIROGA GARCÍA

PROFESOR GUÍA:
CHARLES THRAVES CORTÉS-MONROY

MIEMBROS DE LA COMISIÓN:
ANDRÉS MUSALEM SAID
MARCELO OLIVARES ACUÑA

SANTIAGO DE CHILE
2024

OPTIMIZACIÓN DE LAS ESTRATEGIAS DE PARADA EN BOXES EN LA FÓRMULA 1: UN ENFOQUE BASADO EN DATOS

La Fórmula 1 es conocida como la cumbre del automovilismo, y es la principal categoría de carreras de monoplazas, exhibiendo a los mejores pilotos del mundo y autos de carrera de vanguardia, diseñados para cumplir con rigurosos estándares. La estrategia desempeña un papel fundamental en la Fórmula 1, una buena estrategia puede impulsar a un piloto a ganar posiciones en la pista, evitar rivales y optimizar el ritmo general de la carrera. En este estudio, nos sumergimos en el desafío de la estrategia de neumáticos y paradas en boxes, considerando factores como la degradación, el desgaste y el rendimiento. Logramos esto mediante el desarrollo de un modelo determinista que determina la estrategia de carrera óptima. Es importante destacar que esta optimización se realiza en tiempo real, basándose en estimaciones generadas a partir de datos recopilados solo hasta la vuelta actual.

Nuestra metodología abarca tres fases clave. En primer lugar, empleamos análisis de regresión para evaluar el impacto del desgaste de los neumáticos en los tiempos por vuelta para diferentes compuestos de neumáticos. Dada la presencia de factores de datos no observables que pueden introducir un alto margen de error o sesgos, proponemos dos técnicas de eliminación de valores atípicos. Estos enfoques implican la calibración iterativa de un modelo de regresión y la exclusión de puntos de datos identificados como valores atípicos.

Además, en las primeras etapas de una carrera, cuando el número de puntos de datos (tiempos por vuelta) es limitado, los coeficientes estimados pueden carecer de fiabilidad. Para mitigar esto, calibramos un modelo de regresión previo utilizando datos de carrera del mismo circuito en un año anterior. Al fusionar el modelo previo con los datos de la carrera actual (hasta la vuelta que se está evaluando), obtenemos una distribución posterior. Los coeficientes estimados abarcan varios aspectos, como compuestos de neumáticos, desgaste de neumáticos y pérdidas de tiempo incurridas durante las paradas en boxes. Estos coeficientes se utilizan luego para optimizar la estrategia de carrera.

El proceso de optimización se basa en una formulación de Programación Cuadrática Mixta Entera (MIQP), determinando tanto la cantidad de vueltas como el compuesto de neumáticos para las vueltas que quedan. Es importante destacar que el MIQP se puede resolver eficientemente en una fracción de segundo. Además, proporcionamos un límite superior no trivial para el número máximo de paradas en boxes para el resto de la carrera.

Nuestros resultados numéricos indican que, a partir de la vuelta 20, la estrategia online óptima se alinea estrechamente con la estrategia offline óptima, que implica el uso de estimaciones de regresión basadas en la utilización hipotética de todos los datos de carrera. De acuerdo con estos hallazgos, recomendamos iniciar las carreras con un compuesto medio, lo que permite la acumulación de más datos y mejores estimaciones, evitando los riesgos asociados a comenzar la carrera con un neumático blando.

OPTIMIZATION OF PIT STOP STRATEGIES IN FORMULA 1 RACING: A DATA-DRIVEN APPROACH

Formula 1 is renowned as the pinnacle of motorsport and the premier class of single-seater racing, showcasing the world's top drivers and cutting-edge race cars engineered to meet rigorous standards. Strategy plays a fundamental role in Formula 1, as a well-executed approach can propel a driver to gain track positions, fend off rivals, and optimize the overall race pace. In this study, we delve into the tire and pitstop strategy challenge, considering factors such as degradation, wear, and performance. We achieve this through the development of a deterministic model that determines the optimal race strategy. Notably, this optimization is conducted in real-time, drawing on estimates generated from data collected only up to the current lap.

Our methodology encompasses three key phases. To begin, we employ regression analysis to evaluate the impact of tire wear on lap times for different tire compounds. Given the presence of non-observable data factors that can introduce a high margin of error or biases, we propose two outlier removal techniques. These approaches involve iterative calibration of a regression model and the exclusion of data points identified as outliers.

Furthermore, in the early stages of a race, when the number of data points (lap times) is limited, the estimated coefficients may lack reliability. To mitigate this challenge, we calibrate a prior regression model using race data from the same circuit in a previous year. By merging the prior model with current race data up to the present lap, we obtain a posterior distribution. The estimated coefficients encompass various aspects, including tire compounds, tire wear, and time losses incurred during pit stops. These coefficients are then employed to optimize the race strategy for the remainder of the race.

The optimization process hinges on a Mixed-Integer Quadratic Programming (MIQP) formulation, determining both the length and tire compound for all upcoming stints. Notably, the MIQP can be efficiently solved in a fraction of a second. Additionally, we provide a non-trivial upper limit for the maximum number of remaining pit stops for the race's remainder.

Our numerical results indicate that, commencing from lap 20, the optimal online strategy closely aligns with the optimal offline strategy, which entails using regression estimates based on the hypothetical utilization of all race data. In line with these findings, we recommend initiating races with a Medium compound, allowing for the accumulation of more data and enhanced regression estimates, while avoiding the risks associated with commencing on Soft compounds.

*A los que siempre
confiaron en mi.*

Acknowledgments

Gracias a mi mamá Mariana, a mi papá Juan Pablo, y a mis hermanos, Josefina y Benjamin, quienes han sido un pilar fundamental en mi vida y durante este viaje académico. Les agradezco por su inquebrantable fe en mí, por brindarme las herramientas para alcanzar mis metas y por ser mi inspiración diaria. Además, agradezco a mis tios, mis primas y mi abuela, que han sido mi red de apoyo, compartiendo tanto los momentos de alegría como los desafíos.

A todos mis amigos que han estado a mi lado, celebrando mis éxitos y apoyándome en los desafíos. Gracias por ser una fuente invaluable de apoyo y alegría, y por sus palabras de aliento cuando más las necesitaba.

Finalmente, gracias a mi profesor guía Charles Thraves, quien ha desempeñado un papel fundamental en la culminación de este proyecto de tesis. Siempre dispuesto a compartir sus conocimientos y a brindarme ayuda cuando era necesario.

Este logro no solo es mío, sino también de ustedes. Gracias por ser mi sostén emocional, por celebrar mis triunfos y por ayudarme a superar los obstáculos.

Table of Content

1. Introduction **1**

Literature Review . . . 3

2. Data **5**

2.1. Data . . . 5

2.1.1. Database construction . . . 6

3. Estimation of tire compound performance **8**

3.1. Regression . . . 8

3.2. Outliers detection . . . 13

3.2.1. Interquartile Method . . . 15

3.2.2. Confidence Region Method . . . 16

3.2.3. Comparison of the outliers detection methods . . . 16

3.3. Coefficient priors . . . 23

4. Optimization of race strategy **27**

4.1. Estimate race time of a profile strategy . . . 28

4.2. Optimization of the best running profile . . . 31

4.3. Evaluation of the method for different laps of the race . . . 36

4.3.1. Optimizing race strategies on-line . . . 36

4.3.2. Comparison with offline optimal strategy . . . 39

4.3.3. Analysis of all driver strategies . . . 40

5. Conclusions **44**

Bibliography **46**

ANNEXES **48**

Annex A. Hungarian Grand Prix 2022 **48**

Annex B. Abu Dhabi Grand Prix 2022 **51**

Annex C. Dutch Grand Prix 2022 **59**

Table Index

2.1. Database columns description. 7
3.1. Regression's dependent variables description. 10
3.2. Summary results of the regression considering all laps 11
3.3. Summary results of the regression considering all laps with the *Interquartile* method 19
3.4. Summary results of the regression considering all laps with the *Confidence Region* method 20

Index of Illustrations

2.1. Accumulated data available for each lap at the Hungarian Grand Prix 2022. . 7
3.1. Lap times Hungarian Grand Prix 2022. 9
3.2. Aditional time per lap with each tire for the Hungarian Grand Prix 2022. . . . 12
3.3. Residuals 2022 Hungarian Grand Prix considering all data points. 13
3.4. Comparison of eliminated lap times for the Hungarian Grand Prix 2022. . . . 17
3.5. Comparison of residuals for both methods for the Hungarian Grand Prix 2022. 17
3.6. Comparison of eliminated data for each method for the Hungarian Grand Prix 2022. 18
3.7. Comparison of additional time per lap with each tire for the Hungarian Grand Prix 2022. 21
3.8. Comparison of additional time per lap with each tire at lap 15 for the Hungarian Grand Prix 2022. 22
3.9. Data available for the Hungarian Grand Prix 2022 with the Interquartile method. 24
3.10. Comparison of additional time per lap with each tire for the Hungarian Grand Prix 2022 Interquartile method. 26
4.1. Max Verstappen's stints during Hungarian Grand Prix 2022. 30
4.2. Evolution of the race strategy for Max Verstappen for the Hungarian Grand Prix 2022. 37
4.3. Evolution of the race strategy for Charles Leclerc for the Hungarian Grand Prix 2022. 38
4.4. Evolution of the race strategy for Alexander Albon for the Hungarian Grand Prix 2022. 39
4.5. Comparison of the evolution of race strategy for Charles Leclerc comparing the scenario where the prior race is taken into account and where it is not considered. 40
4.6. Time difference between our proposed strategy and the optimal strategy for Albon, Leclerc, and Verstappen during the Hungarian Grand Prix 2022, evaluating every 5 laps. 41
4.7. Time difference between our proposed strategy and the optimal strategy for every driver during the Hungarian Grand Prix 2022, evaluating every 5 laps. . 41
4.8. Time difference between the real strategy used by the driver and the optimal strategy we propose for every driver during the Hungarian Grand Prix 2022. . 42
4.9. Time difference between the real strategy used by the driver and the optimal strategy we propose for every driver during the Hungarian Grand Prix 2022, considering the team effect. 43
A.1. Comparison of the evolution of race strategy for Max Verstappen comparing the scenario where the prior race is taken into account and where it is not considered. 48
A.2. Comparison of the evolution of race strategy for Alexander Albon comparing the scenario where the prior race is taken into account and where it is not considered. 49

A.3. Time difference between the real strategy used by the driver and the optimal strategy we propose for every driver during the Hungarian Grand Prix 2022. . 50
B.1. Lap times Abu Dhabi Grand Prix 2022 51
B.2. Comparison of the tire performance at lap 15 Abu Dhabi Grand Prix 2022. . . 52
B.3. Tire performance at the end of the Abu Dhabi Grand Prix 2022, considering data from the previous year's race 52
B.4. Evolution of the race strategy for Max Verstappen for the Abu Dhabi Grand Prix 2022. 53
B.5. Comparison of the evolution of race strategy for Max Verstappen comparing the scenario where the prior race is taken into account and where it is not considered for the Abu Dhabi Grand Prix 2022. 53
B.6. Comparison of the evolution of race strategy for Yuki Tsunoda comparing the scenario where the prior race is taken into account and where it is not considered for the Abu Dhabi Grand Prix 2022. 54
B.7. Comparison of the evolution of race strategy for Mick Schumacher comparing the scenario where the prior race is taken into account and where it is not considered for the Abu Dhabi Grand Prix 2022. 55
B.8. Time difference between our proposed strategy and the optimal strategy for every driver during the Abu Dhabi Grand Prix 2022, evaluating every 5 laps. . 56
B.9. Time difference between the real strategy used by the driver and the optimal strategy we propose for every driver during the Abu Dhabi Grand Prix 2022. . 57
B.10. Time difference between the real strategy used by the driver and the optimal strategy we propose for every driver during the Abu Dhabi Grand Prix 2022. . 58
B.11. Time difference between the real strategy used by the driver and the optimal strategy we propose for every driver during the Abu Dhabi Grand Prix 2022. . 58
C.1. Lap times Dutch Grand Prix 2022 59
C.2. Comparison of the tire performance at lap 15 Dutch Grand Prix 2022. 60
C.3. Tire performance at the end of the Dutch Grand Prix 2022, considering data from the previous year's race 60
C.4. Evolution of the race strategy for Max Verstappen for the Dutch Grand Prix 2022. 61
C.5. Comparison of the evolution of race strategy for Max Verstappen comparing the scenario where the prior race is taken into account and where it is not considered for the Dutch Grand Prix 2022. 61
C.6. Evolution of the race strategy for Lewis Hamilton for the Dutch Grand Prix 2022. 62
C.7. Comparison of the evolution of race strategy for Lewis Hamilton comparing the scenario where the prior race is taken into account and where it is not considered for the Dutch Grand Prix 2022. 63
C.8. Time difference between our proposed strategy and the optimal strategy during the Dutch Grand Prix 2022, evaluating every 5 laps. 64
C.9. Time difference between the real strategy used by the driver and the optimal strategy we propose for every driver during the Dutch Grand Prix 2022. 65
C.10. Time difference between the real strategy used by the driver and the optimal strategy we propose for every driver during the Dutch Grand Prix 2022. 66
C.11. Time difference between the real strategy used by the driver and the optimal strategy we propose for every driver during the Dutch Grand Prix 2022. 66

Chapter 1

Introduction

Formula 1 is the pinnacle of motorsport and the highest class of single-seater racing. It features the best drivers in the world and high-tech race cars that are built to meet several requirements. Approximately 80,000 components come together to make an F1 car, and they are designed to reach more than 1,000 horsepower. These extremely aerodynamic cars contain turbocharged engines and hybrid systems that gather heat energy throughout the race and allow short bursts of extreme speed. The question now is, what's the main objective of competing in Formula 1? To win the championship. To accomplish this, drivers have to win races.[1] Winning in F1 depends not only on the driver's performance but also on the efficiency and expertise of their team. Bell et al. (2016) conducted a comprehensive multilevel modeling study of Formula One Driver and Constructor performance, showing that "team effects significantly outweigh driver effects, accounting for 86% of driver variation".

Winning a race combines many factors, such as the driver's skills, the weather, the track, the competitors, the pit stop strategy, and many other events that may happen during a race. Formula 1 races are not just about speed; they are also about managing resources and making the right choices at the right time. Strategy plays a fundamental role in Formula 1. A well-executed strategy can help a driver gain positions on the track, defend against competitors, and maximize their overall race pace. This involves considering factors such as tire degradation, track conditions, weather forecasts, and the positions of other drivers.

Formula One season consists of a series of races, known as Grand Prix. A Grand Prix takes place in multiple countries and continents around the world on either purpose-built circuits or closed public roads and they are held over a weekend. A typical Grand Prix weekend[2]

[1] Formula 1 gives points to the first 10 classified in the race. At the end of the season, the driver with the most points wins the championship. As of the 2010 season, the system allows scoring the top 10, with a more progressive scale than the previous one, to encourage the fight for leading positions. It awards 25 points to the first classified, 18 to the second, 15 to the third, 12 to the fourth, 10 to the fifth, 8 to the sixth, 6 to the seventh, 4 to the eighth, 2 to the ninth, and 1 to the tenth classified. In addition, from 2019, a point will be awarded to the driver among the top ten classified who achieves the fastest lap, losing the effect when the driver who achieves the fastest lap is located in eleventh place or lower.1

[2] There is a new format called "Sprint Weekend" format, which refers to a new race weekend format that was introduced to a few Grand Prix weekends in the 2021 season. In this format, the drivers will only run one free practice session over the weekend, which will be on Friday morning. On Friday afternoon the classic classification will be run, which will determine the starting grid on Sunday. On the other hand, on Saturday it will be the turn of the Sprint Shootout. This sprint classification set the grid for the 100-kilometer sprint race that also takes place on Saturday, called the Sprint Race.

normally begins on Fridays, when Practice sessions are held. Qualifying sessions are held on Saturdays and will determine the race's starting order. The Grand Prix is held on Sunday and is the weekend's main event. Under normal circumstances, the winner of the race is the first driver to cross the finish line having completed a set number of laps.

Pirelli is the official Formula 1 tire supplier. Each year, they develop different types of tires for the teams to choose from. They provide three different compounds of the slick tire, as well as the intermediate and full-wet tire, for teams to use over a race weekend. There are six slick compounds within the range, numbered from zero to five from hardest to softest. These are known as C0 to C5, with the 'C' standing for 'compound'. From this range of tires, only three are available per weekend, which are denoted as Hard, Medium, and Soft. Moreover, there are intermediate and full wets tires to choose from for wet conditions. Depending on the track surface, temperature, humidity, and other conditions, teams will adjust their strategies accordingly to maximize performance. On any given weekend, the way teams use each compound differs considerably.

During a race, a driver is required to use at least two different types of tires. On each lap, each car faces the decision of either making a pitstop or continuing on track. This decision has to do with the rule mentioned above, and with tire degradation. The Soft tire is the softest compound in the range. It is designed for the slowest circuits with low wear and degradation where maximum mechanical grip is required from the rubber. This tire is the fastest over one lap but also has the highest level of wear. These tires are normally seen at street circuits or where the asphalt is exceptionally smooth. The Medium tires can be recognized by the yellow stripes. With an excellent balance between performance and durability, this tire is well-suited to a wide range of conditions. This is often viewed as the strongest tire for a race due to its often long lifespan and a considerable pace advantage over the hard equivalent. Finally, we have the hardest tire in the range, we can recognize it by the color white. It is designated to provide maximum resistance to heat and extreme forces, lasting longer than the other compounds. This tire is capable of running very long stints, but it comes at the expense of performance. Lacking the pace of the soft option, this compound puts a driver at a disadvantage over a single lap or a short stint but can yield benefits across a longer period. For wet conditions, there are intermediate and full wets to choose from. The intermediates are the most versatile type of tire when it comes to racing in wet conditions. They can be used on a wet track with no standing water, as well as a drying surface. The compound has been designed to have a wide working range, guaranteeing a wide crossover window both with the slicks and the full wets. They provide superior performance until the track is ready for slick tires. The full wet tires are the most effective for heavy rain, capable of dispersing impressive quantities of water. These tires should only be used when there is standing water on the track as they are ineffective in a drying track environment.

In Formula 1, tire strategy plays a crucial role in the success of a race. Teams must choose the right tires for each track to maximize performance and minimize risks of failure. The right tire compounds and strategies can help a team gain an advantage over its competitors. Tire compounds are chosen based on the track conditions, driver style, and the team's strategy for a particular race. Many factors can affect the strategy. One of them is the trade-off between using a softer compound for maximum speed but with high degradation or using a harder compound with less performance but long-lasting. Teams must also manage tire wear

carefully, by monitoring the pressure and temperature of the tires. They also need to check for any wear or damage. During the race, the car gets lighter because of the fuel expense. This means that the degradation of the tires decreases and the performance improves. Ultimately, the right tire strategy can help a team gain an edge over its competition and come out on top.
Another factor that plays a crucial role in a team's strategy has to do with the complexity of the problem when solving the optimization online, that is, when there is only partial information. Meaning that we can only optimize with the information that is available, from the start of the race to the current lap.

In this work, we address the tire and pitstop strategy problem considering some of the issues mentioned above. Specifically, we would work with degradation, wear, and performance, by formulating a deterministic model to determine the optimal race strategy. For this, we consider each lap as a different stage in the race, where there would be a decision to continue with the same tires or make a pitstop. In this model, we would consider the following variables to describe the current state of each car: current lap, current tire, lap times, compounds used, tire wear, teams, and the interaction with other drivers by using the gaps between them and considering if they are lapped or racing cars. Even though we would not be working with yellow flags, or considering the weather, to make it close to reality we would optimize the race strategy considering all the available data at the moment of the decision. This means that if we are on lap 20, we would use all the data until lap 19 of all drivers. And if we are at the beginning of the race, we would use the data of the year before at the same location. This would update the model, to consider the new conditions of the driver and the track.

Literature Review

The FIA Formula One World Championship has been one of the premier forms of racing worldwide since its inaugural season in 1950, with Formula 1 (F1) attracting extensive research, particularly in the realm of aerodynamics. Agathangelou and Gayscone (1998) [1] quantified the significance of aerodynamics in modern F1 cars and explored the impact of FIA regulations on aerodynamic development. Recent works, such as Tianyou Hu's (2023) [16], delve into Venturi Wind Tunnel and ground effects in F1 and aircraft, explaining their substantial influence.

In the domain of race strategy optimization, Bekker and Lotz (2009) [3] employed discrete-event simulation techniques to plan and analyze F1 race strategies. Their work focuses on enhancing decision-making through simulation methods. In 2013, Perantoni and Limebeer [18] tackled the minimum-lap-time optimal control problem for F1 race cars using direct transcription and nonlinear programming. Tremlett and Limebeer (2015) [25] explored tire utilization strategies to improve performance. Regarding tire management, West and Limebeer (2020) [27] extended a thermodynamic model to include tire carcass temperature, employing optimal control calculations to enhance grip and tire longevity.

Heilmeyer, Graf, and Lienkamp (2019) [13] initiated race simulations considering various race strategy inputs, including tire degradation, fuel mass loss, pit stops, and overtaking maneuvers. Their lap-wise discretization approach incorporates publicly accessible lap time data. Building on this, Heilmeyer, Thomaser, Graf, and Betz (2020) [14] leveraged neural networks to create a "Virtual-Strategy Engineer" capable of predicting optimal race strategies based on historical race data, enhancing decision-making during races.

Talbi (2015) [24] explored optimization techniques for hairpin turns during races, modeling 180-degree turns as network flows on a grid. Real-time decision-making in motorsports was addressed by Tulabandhula and Rudin (2014) [26], focusing on predictive analytics for factors like tire changes, track conditions, and yellow flags. Choo (2015) [6] emphasized real-time decision-making in motorsports to enhance strategic decisions made by racing teams.

Sarabakha, Pichler, and Jacak (2015) [12] investigated the use of simulation, machine learning, and optimization, presenting a case in Formula 1 (F1) competition with a decision support system (DSS) framework. Duhr, Buccheri, Balerna, Cerofolini, and Onder (2023) [8] addressed energy management optimization in F1 race cars with hybrid-electric powertrains, focusing on setting lap-by-lap targets for fuel and battery consumption. In the context of lap times, Borsboom, Fahdzyana, Hofman, and Salazar (2021) [5] introduced a convex optimization framework for minimum lap time design and control of electric race cars, presenting a systematic approach for calculating minimum-lap-time control strategies.

In the most recent research from 2023, Heine and Thraves (2023) [15] proposed a systematic and mathematical approach to optimizing pit-stop strategies in motorsport, considering various variables and uncertainties affecting race outcomes.

Overall, these works contribute significantly to understanding and enhancing various aspects of Formula 1 racing, from aerodynamics and race strategy to lap time optimization and real-time decision-making. The key distinction lies in our approach, solving the optimization online by incorporating data from the previous year's race as a prior belief.

Chapter 2

Data

This chapter explains the origin of the data, the available data, and exploratory data analysis to gain greater clarity on the structure and the quality of the data available.

2.1. Data

The dataset "Formula 1 World Championship (1950-2020)" by Rao, Rohan[21], was obtained from Kaggle. It contains information on the Formula 1 races, drivers, constructors, qualifying, circuits, lap times, pit stops, and championships from 1950 till the latest 2023 season. The dataset includes a wide range of information, such as:

- Race Data:
  - Race details: Race name, date, location, circuit, round number, and race ID.
  - Race results: Finishing positions, starting positions, and race times for each driver.
  - Fastest laps: Fastest lap time and lap number for each driver.
  - Constructor points: Points earned by each constructor in the race.
- Driver and Constructor Information:
  - Driver Details: Driver names, nationality, code, and driver ID.
  - Constructor details: Constructor names, nationality, and constructor ID.
- Qualifying Data:
  - Qualifying results: Qualifying positions and times for each driver.
  - Qualifying weather conditions.
- Lap Times:
  - Lap times for each driver in each race. It contains Race ID, Driver ID, race lap, position, and lap time in milliseconds.
- Pit Stops:
  - Pit stop information: Lap number of each pit stop, pit duration, and tire change details.

- Circuits:
  - Information about each racing circuit.

Besides this, to obtain information to build the data set about the stints, it was used the website Racefans[19]. It contains information about which tire each driver used during the race and how many laps they were used. For every race, we have a database with all the pitstops during that race. It has the driver's name, and the stints, with the compound and the number of laps the driver used every tire.

2.1.1. Database construction

From all the available data, we use the following databases.

- Race details: Race name, date, location, circuit, round number, and race ID.
- Driver details: Driver names, nationality, code, and driver ID.
- Pit stop information: Lap number of each pit stop, pit duration, and tire change details.
- Stints: Driver's code, and the stints, with the compound and the number of laps the driver used every tire.
- Lap times for each driver in each race. It contains Race ID, Driver ID, race lap, position, and lap time in milliseconds.

In addition to the variables provided in the described datasets, we construct additional variables that impact a car's lap time. These variables are the following:

- Time difference to the first car not lapped[3] behind.
- Time difference to the first car not lapped in front.
- Time difference to the first car lapped in front.
- DRS[4]

It's important to note that we make a difference if the car in front is lapped or not. This is because of the effect of blue flags. When a faster car is approaching a lapped car from behind, race marshals display blue flags to inform the lapped driver that a faster car is approaching and that they should move over to allow the faster car to pass. But, if the car in front is not lapped, there will not be a blue flag, meaning that they will fight for position.

Thus, as seen in Table 2.1, the database that we will be working with, contains the following columns:

[3] A "lapped car" refers to a competitor whose position on the track is one or more laps behind the race leader.

[4] DRS stands for "Drag Reduction System," and it is a technology used to enhance overtaking opportunities and improve the racing spectacle. On certain sections of the race track, known as "DRS Zones," there are sensors that monitor the distance between cars. When a pursuing car gets within a 1-second gap of the car in front, the DRS system becomes available for that pursuing car. This opens a moveable flap on the car's rear wing, reducing the aerodynamic drag and allowing the pursuing car to achieve higher top speeds on straights.

Table 2.1: Database columns description.

Column name	Description
Lap	Lap of the race
Position	Position of the driver
Lap Time	Lap time
Accumulated Lap Time	Accumulated lap time
Driver	Driver's code. Each driver has a unique code that represents them during the race. It has three letters, and they usually are their first three last name letters.
Team	Driver's team
Tire Compound	Tire compound
Lapwear	Lapwear of the current tire they are using
Inpits	Binary variable that will be 1 if the driver is pitting on the lap being evaluated, and 0 if not
Outpits	Binary variable that will be 1 if the driver is leaving pits on the lap being evaluated, and 0 if not
DRS	Binary variable that will be 1 if the driver has DRS, and 0 if not
Δ_1	Time difference to the first car not lapped behind
Δ_2	Time difference to the first car not lapped in front
Δ_3	Time difference to the first car lapped in front

Despite we have data from multiple races, the proposed methodology is going to be applied within the same race. More specifically, the goal is to optimize a driver pit stop strategy during a race by only considering the available data from the previous laps of the same race (including other drivers as well). This data is used to estimate the tire compound performance (in terms of the impact of tire wear on a car's lap time). The reason why we do not use data from other circuits is that races are very heterogeneous concerning lap times, tire degradation, and specific tire compounds[5], among others. In this study, we would be working with the 2022 Hungarian Grand Prix because it was a race where some teams had trouble finding the right tire strategy.

In this work we seek to make estimates in each of the laps, therefore, in each lap of the race, there will be more data on which to calibrate the predictive models. Figure 2.1 shows the accumulated data available (y-axis) in each lap of the race (x-axis).

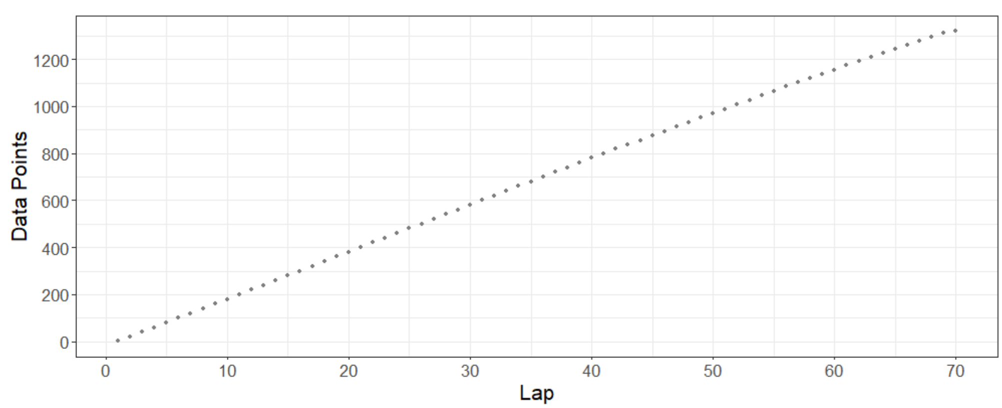

Figure 2.1: Accumulated data available for each lap at the Hungarian Grand Prix 2022.

As we can see, as the race progresses, there is more data available to calibrate the models. Thus, we would be able to make better predictions of tire compound performance.

[5] For each race there are three tire compounds chosen from a total of six, which might not match from one race to another one.

Chapter 3

Estimation of tire compound performance

Before going into the optimization of race strategies, we look into the problem of estimating the impact of tire wear in lap times for all tire compounds. A way to perform this is by using regression analysis.

3.1. Regression

Regression analysis is a statistical process that allows us to analyze the relationship between two or more variables, one dependent on the rest of the variables. By doing these, we can estimate the performance[6] of each tire compound, in terms of lap time, and uses it to optimize the tire strategy for the remainder of the race.
Since we want to analyze the tire compound performance, we want to work with a race where teams had trouble finding the right strategy, for example, the Hungarian Grand Prix in 2022. Nevertheless, this work can be easily used in any other race.

First, let's look at the lap times for each driver during the race. Figure 3.1 displays all drivers' lap times (y-axis) for each lap during the race (x-axis). As we can see, we have some lap times that are higher than the rest, most of them can be explained by 3 main reasons. On the first lap, cars are grouped and fighting for position, which can create traffic and make overtaking difficult. Also, at the start of the race, the tires are relatively cold, which reduces their grip on the track. Drivers can find themselves fighting for positions and avoiding collisions. Another explanation would be pit stops since they have to slow down to enter the pitlane. After completing the pit stop, drivers must re-enter the track from the pit lane. This involves a process of acceleration and merging into normal race traffic that takes extra time. Finally, we have yellow flags[7]due to an incident on the race track, where drivers must reduce their speed, and therefore, increase their lap times significantly.
In addition, we can appreciate that, on one hand, as the race progresses, the lap times are progressively decreasing. This is because the fuel load is lower, making the cars lighter, and

[6] The performance of tires in F1 refers to how well they behave on the racing circuit. It considers the grip on the track surface, how quickly they wear down, and how they respond to various track conditions.

[7] The yellow flag is a crucial signal used to indicate a potentially dangerous situation on the track. When a yellow flag is displayed, it warns drivers to slow down, and be prepared to encounter a hazardous situation ahead, such as a car that has spun off the track, debris on the circuit, or track marshals on the track.

therefore faster. On the other hand, we can see that as a tire is used over more laps, there is a corresponding increase in lap times. Thus, there is a trade-off between making or not making a pit stop. Although making a stop takes time, it allows them to have a fresh tire, and improve their lap times.

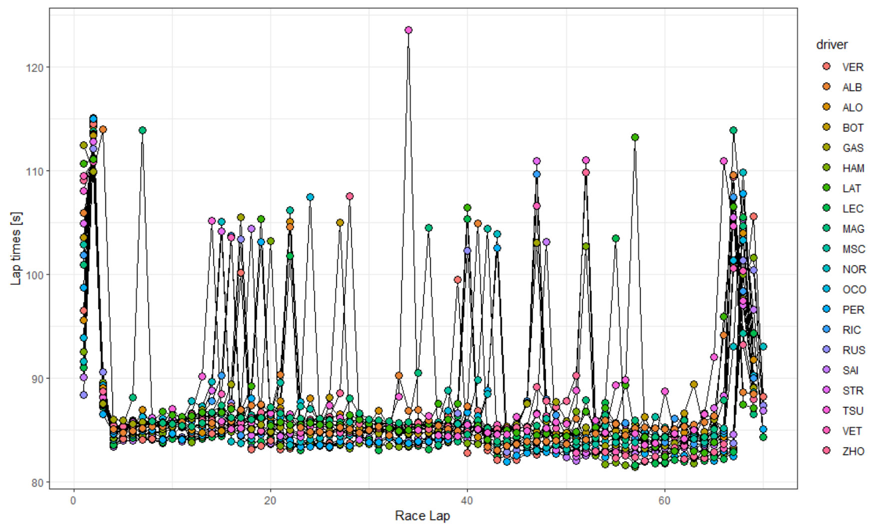

Figure 3.1: Lap times Hungarian Grand Prix 2022.

Let us denote by N the total number of laps of a race, equal to 70 in the case of the Hungarian Grand Prix. The set of drivers is denoted by $\mathcal{D}$ and the set of constructors teams is denoted by $\mathcal{C}$. In particular, these sets correspond to the following:

- $n \in \{1, \ldots, N\}$.
- $\mathcal{D} \in \begin{Bmatrix} \text{Ver, Per, Ham, Rus, Lec, Sai, Vet, Str, Alo, Oco, Tsu, Gas, Msc, Mag, Nor,} \\ \text{Ric, Zho, Bot, Alb, Lat} \end{Bmatrix}$
- $\mathcal{C} \in \begin{Bmatrix} \text{Red Bull, Ferrari, Mercedes, Aston Martin, Alpha Tauri,} \\ \text{Haas, McLaren, Alfa Romeo, Williams, Alpine} \end{Bmatrix}$

From the above, we define the variable y_{dn}. The lap time in seconds for driver d during lap n. Where $d \in \mathcal{D}$ and $n \in N$.

Through regression, we can examine how changes in the independent variables affect the dependent variable so we can predict future or unknown values. In this work, we will be using it to study the effect of wear on the different types of tires. This methodology allows us to understand patterns and trends in the data, and in the end, find the best tire strategy for the race. Although we are interested in estimating the impact of tire wear for different compounds in the lap times, there are several other variables we need to control for in the regression. A list of these variables, including their notation and description is provided in Table 3.1. Then, the regression equation can be expressed as:

$$y = \beta_0 + \beta_1 X_{lap} + \beta_2 X_{lap}^2 + \beta_3 X_{inpits} + \beta_4 X_{outpits} + \sum_{c \in \mathcal{C}} \beta_c X_c + \beta_{\Delta 1} \cdot \frac{1}{1+X_{\Delta_1}} + \beta_{\Delta 2} \cdot \frac{1}{1+X_{\Delta_2}} + \beta_{\Delta 3} \cdot \frac{1}{1+X_{\Delta_3}} + X_{s0}(\beta_0^s + \beta_1^s X_w) + X_{0m}(\beta_0^m + \beta_1^m X_w) + X_{h0}(\beta_0^h + \beta_1^h X_w)$$

For simplicity, we have removed the indexes of the driver-lap tuple, (d, n), in the dependent and independent variables of the regression in Equation 3.1. Where we have that $\Delta 1$, $\Delta 2$ and $\Delta 3$, refer to the time difference to the first car not lapped behind, the time difference to the first car not lapped in front, and the time difference to the first car lapped in front. All of them in seconds. Then, the dependent variables are the following:

Table 3.1: Regression's dependent variables description.

Variable	**Values**	**Description**
X_{lap}	$\{1, \ldots, 70\}$	Race lap.
X_{inpits}	$\{0, 1\}$	Binary variable that will have a value of 1 if the driver is pitting on the lap being evaluated, and 0 if not.
X_{inpits}	$\{0, 1\}$	Binary variable that will have a value of 1 if the driver is pitting on the lap being evaluated, and 0 if not.
$X_{constructor}$	{Red Bull, Mercedes, Ferrari, Aston Martin, Alpine, Alpha Tauri, Haas, McLaren, Alfa Romeo, Williams }	Qualitative variable that returns the team of the pilot.
$X_{\Delta 1}$	$[0, 1.5]$	Time difference to the first car not lapped behind.
$X_{\Delta 2}$	$[0, 1.5]$	Time difference to the first car not lapped in front.
$X_{\Delta 3}$	$[0, 1.5]$	Time difference with the first car lapped in front.
X_w	$\{1, \ldots, 70\}$	Number of laps a type of tire is used during a stint.
X_{s0}	$\{0, 1\}$	Binary variable that will have a value of 1 if the car is going to use a soft tire, and 0 otherwise.
X_{m0}	$\{0, 1\}$	Binary variable that will have a value of 1 if the car is going to use a medium tire, and 0 otherwise.
X_{h0}	$\{0, 1\}$	Binary variable that will have a value of 1 if the car is going to use a hard tire, and 0 otherwise.

As we can see in the table above, the time difference to the first car not lapped behind, the time difference to the first car not lapped in front, and the time difference to the first car lapped in front can take values between 0 and 1.5. This upper limit is given because if the car is at a distance greater than 1.5 seconds, we assume that the effect on the car will not be significant. We are considering that if the car is less than 1 second away DRS can be used, and we give a margin of 0.5 seconds extra where the battle between the cars affects performance.

We run the regression for the particular case in which we have all the laps of the race. Note that this could not be used during the on-line setting in which we ran the regression for some laps in the middle of the race. Table 3.2 shows the summary results of the regression considering all laps.

Table 3.2: Summary results of the regression considering all laps

term	estimate	std.error	statistic	p.value
(Intercept)	92.56	0.69	134.09	0.00
lap	-0.56	0.03	-20.34	0.00
lap^2	0.01	0.00	19.18	0.00
inpits	4.01	0.71	5.65	0.00
outpits	21.24	0.71	29.74	0.00
escuderiaAlfa Romeo	1.29	0.57	2.26	0.02
escuderiaAlpha Tauri	2.71	0.54	5.01	0.00
escuderiaAlpine	1.07	0.62	1.73	0.08
escuderiaAston Martin	1.73	0.54	3.21	0.00
escuderiaFerrari	-0.08	0.52	-0.15	0.88
escuderiaHass	1.17	0.55	2.11	0.04
escuderiaMcLaren	0.85	0.55	1.55	0.12
escuderiaMercedes	-0.10	0.52	-0.20	0.84
escuderiaWilliams	2.17	0.53	4.07	0.00
fungapsecatrasnorezagado	-1.21	0.64	-1.90	0.06
fungapsecadelantenorezagado	-1.29	0.65	-1.98	0.05
fungapsecadelanterezagado	-0.88	1.19	-0.74	0.46
Medium	-0.06	0.52	-0.12	0.91
Hard	2.18	0.70	3.11	0.00
Soft·lapwear	-0.02	0.04	-0.58	0.56
Medium·lapwear	0.06	0.02	2.55	0.01
Hard·lapwear	-0.03	0.03	-1.09	0.28

The coefficient of determination R^2 obtained is 0.501, with a total of 1298 data used.

Based on the p-values, we observe that the variables Lap, Inpits, Outpits, Hard, and some teams are statistically significant predictors of the lap time. The time differences, and the lineal components of the tires, on the other hand, do not show a statistically significant impact on race time.
If we take a look at the team performances, we get that the fastest team is Mercedes, followed by Red Bull, and Ferrari, with a difference of 0.01 seconds by lap between them. We also can see that the slower teams on the grid are Williams and Alpha Tauri, with more than a 2-second gap to the fastest teams. Thus, we see that the team has a big effect on the lap time.
On the other hand, the effect of a pit stop can be observed through the Inpits variable. Thus, one pit stop adds 21.24 seconds to the lap time. In addition, the effect of the fuel can be seen in the Lap variable, since as the race progresses the lap times decrease, showing that as the cars have less fuel, they are lighter, and the lap times decrease.
Finally, we can obtain information about the tire's performance. We have that the Medium tire is the fastest, then we have the Soft, 0.06 seconds slower per lap, and finally the Hard, 2.24 seconds slower per lap. Also, when looking at the slopes, we see that the the Soft and

the Hard tire has a negative slope, showing that as more laps is being used, the better its performance. Then, we see that the slope for the Medium tires is positive.

To continue with the estimation of the tire performance, from the coefficients obtained in the regression, taking a new Soft tire as a reference, we can plot the performance of all tire compounds, obtaining the additional time per lap in comparison.

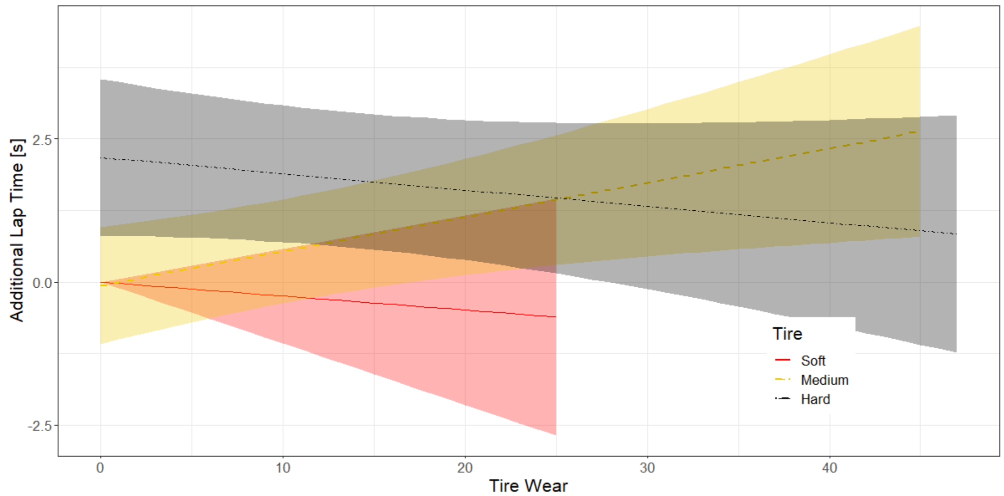

Figure 3.2: Aditional time per lap with each tire for the Hungarian Grand Prix 2022.

In Figure 3.2 we can see that the results are not 100% accurate. As shown, the slopes of the Hard and Soft tires are almost the same, contradicting reality, as the slopes should not be negative. Both slopes should cross at some point, which is not the case. Furthermore, we can see that although the behavior of the Medium compound is close to reality, it is not realistic, since its intercept should not be negative. These results can be explained because there are outliers that don't have an explanation. As seen above, many outliers can significantly impact data analysis and modeling.

Furthermore, it is important to note that we evaluate the performance of the regression. For this, we will calculate the residuals. The residuals of a regression analysis are the differences between the observed values of the dataset and the estimated values calculated with the regression equation. These allow us to know the error in the estimate for the lap times.

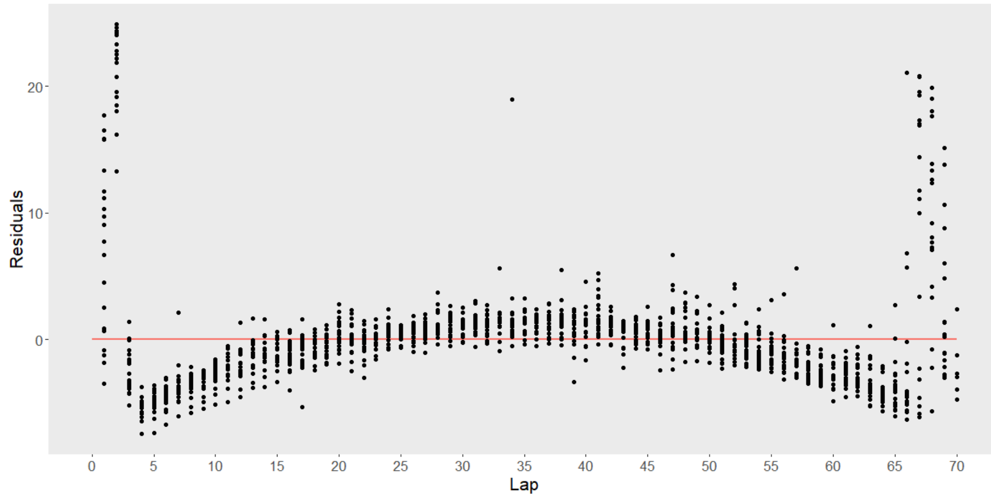

Figure 3.3: Residuals 2022 Hungarian Grand Prix considering all data points.

As we can see in Figure 3.3, we have data points that significantly deviate from the rest of the data. They are observations that are located far away from the other data points and can potentially impact the results of our analysis. These residuals show some big differences between the observed values and the predicted values on some of the values obtained from the regression model, especially in the first 10 laps of the race, and in the last 10. This could be explained because during the first laps of the race, there are usually more incidents, and the cars are adapting to the track conditions. During the latter, the tires tend to be more worn than normal, and with less fuel, performance tends to be more uncertain. It is also important that it can be observed that the concave shape of the residuals can be explained by the fact that, given the quadratic nature of $\hat{y}$, which is convex, and since the residuals are calculated as $y_i - \hat{y}_i$, this value becomes negative.

3.2. Outliers detection

As it was explained previously, there are non-observable factors in the data that generate predictions with high error. But, we have to make a difference between the data that we can explain to, and those that we can not. We have three main reasons to explain these residuals. Large residuals, in terms of absolute value, are because of elements we are not controlling in the independent variables. Some of these non-observable factors could be the following:

- Yellow flags events. We can see that in the first lap and in the last 3 laps of the race, the times were higher than normal, this is because a Virtual Safety Car(VSC)[8] was deployed.

[8] The Virtual Safety Car (VSC) system is a safety measure introduced to manage potentially hazardous situations on the track without the need to deploy a physical safety car. When activated by race officials due to incidents such as debris on the track or a stranded car in a dangerous location, the VSC imposes a controlled reduction in the speed of all cars on the track. This is achieved through a standardized speed delta that drivers must stick to, maintaining a safe and consistent gap to the car in front. By avoiding the deployment of a physical safety car, which gathers the entire field behind it and can lead to significant changes in race dynamics, the VSC aims to create a fairer and less disruptive solution. During a VSC period, drivers are not allowed to overtake, promoting controlled and safe driving.

Pilots must reduce their speed during a VSC and maintain a constant predetermined speed based on VSC signaling. This speed is significantly lower than normal running speed. As a result, lap times are higher due to the limit on the speed drivers can drive. The VSC, not also affects the laps when it was deployed, but also a few laps after it as well. Once the VSC is lifted and the drivers return to normal racing speed, it may take time for the drivers to regain their pace and adjust again to normal racing conditions. This transition period can influence lap times immediately after the VSC is removed.
It is important to note that this effect can not be added to a new variable because of the heterogeneous effect it has on every driver. Let's say that there was a Yellow flag during the first lap of the race. The lap times of the first lap for the drivers that were on the first positions, were not highly affected because they were already finishing that lap, so the effect was on lap 2. On the contrary, for the drivers in the last positions, the majority of the effect was on lap 1. This difference makes it very hard to capture the effect of Yellow flags on a variable.

- First lap. In Formula 1, lap times for the first lap of a race are often higher due to several factors:
  - Cold tires: At the start of the race, the tires are relatively cold, which reduces their grip on the track. Cold tires can affect the drivers' ability to corner at high speeds and brake effectively, resulting in slower lap times.
  - Fuel at full capacity: At the start of the race, the cars are loaded with a large amount of fuel, as they must complete the entire distance of the race. The added weight of the fuel affects the car's performance and acceleration, which is reflected in higher lap times.
  - Traffic and fights for position: On the first lap, cars are grouped and fighting for position, which can create traffic and make overtaking difficult.
  - Conservative Strategy: On the first lap, drivers tend to be more cautious and conservative as they are trying to avoid accidents or damage to their cars. This can lead to a more careful approach instead of going for fast lap times.

  In the race we are currently working with, the Hungarian Grand Prix 2022, it is very difficult to capture these effects because there was a Yellow flag during the first lap of the race.

- Slow Pit Stops. Technical problems with the equipment, such as the wheel nut not being properly secured, malfunctioning jacks, or issues with the equipment used to remove and attach tires, can lead to slow pit stops. Another reason could be a potential safety issue, such as a loose tire that could pose a danger to others on the track, the team might take extra time to ensure safety before sending the car back out.
- Unexpected maneuver. Due to the high speeds and complex aerodynamics of the cars, they can spin. Meaning that the car loses traction and spins around, often completing a full 360-degree rotation.
- Cars with different race paces. If a car is behind a slower one, with a high difference in its pace, it can cause the car behind to have an unusual lap time.

We believe that the counter-intuitive results found (concerning tire compounds performance, see Figure 3.2) in the regression when considering all data points (i.e., lap times) are because of the presence of outliers. We propose two methodologies to remove outliers. They iterate by calibrating a regression and removing data points whose residuals are labeled as outliers. Details are described in the following sections.

3.2.1. Interquartile Method

This method is based on percentiles and the distribution of the data. For this method, the residuals are calculated from the regression. First, we find the 25th percentile (Q_1) and the 75th percentile (Q_3) of the residuals. The Interquartile Range (IQR) method eliminates outliers by identifying data points that fall below $Q_1 - k \cdot IQR$ or above $Q_3 + k \cdot IQR$. Where IQR is expressed as IQR=$Q_3 - Q_1$. Also, we would use $k = 3$, the value usually used in this algorithm. Now we are ready to identify the outliers.
First, we define the residuals.

$$\hat{e_i} = y_i - \hat{y_i}$$

With,

- $\hat{y_i}$= Predicted lap time for lap i with the regression.
- y_i = Observed lap time for lap i.

The method eliminates rows as follows: those that met any of the following conditions would be eliminated.

- $\hat{e_i} < Q_1 - k \cdot IQR$
- $\hat{e_i} > Q_3 + k \cdot IQR$

All values that met these conditions are considered outliers and are marked down, for further removal. It is important to notice that this method does not eliminate the outliers of the lap times, but eliminates those observations whose prediction is not accurate. Meaning that there is no known explanation for the reason for the higher lap time. This algorithm iterates until it does not eliminate any more outliers. We can see the algorithm pseudo-code shown in *Algorithm* 1.

Algorithm 1 Interquartile range algorithm

Require: $I = \{1, \ldots, N\}, k = 3$
  while true **do**
    $\hat{y} \leftarrow LinReg\left(\left\{\left(x^{(i)}, y^{(i)}\right)\right\}_{i \in I}\right)$
    $\hat{e_i} \leftarrow y_i - \hat{y_i} \quad \forall i \in I$
    $Q_3 \leftarrow 75\%$ percentile of $\{\hat{e_i}\}_{i \in I}$
    $Q_1 \leftarrow 25\%$ percentile of $\{\hat{e_i}\}_{i \in I}$
    $IQR \leftarrow Q_3 - Q_1$
    $J \leftarrow \{i \in I | \hat{e_i} \notin [Q_1 - k \cdot IQR, Q_3 + k \cdot IQR]\}$
    $I \leftarrow I \setminus J$
    if $|J| = 0$ **then**
      break
    end if
  end while

3.2.2. Confidence Region Method

The Confidence Region-based outlier removal method is another technique used to identify and remove outliers in our data set. This method is based on the $\hat{y}_i$ obtained from the regression, the goal is to be able to compare it with the y_i observed. First, we obtained $\hat{y}_i$, predicted with the regression, along with the standard deviation of the prediction, s_i.

The idea is to compute for each data the probability of obtaining an outcome as odd as the one obtained according to the regression. Namely, a p-value. Thus, if $y_i < \hat{y}_i$ this probability is computed as $2 \cdot \Phi(\frac{(y_i - \hat{y}_i)}{s_i})$. Otherwise, we will use $1 - 2 \cdot \Phi(\frac{(y_i - \hat{y}_i)}{s_i})$. It's important to note that this fraction follows a standard normal distribution.

Based on the above, the criteria for eliminating rows will be the following: those rows whose previously calculated probability value is less than a certain $\alpha = 10^{-20}$ will not be considered. This algorithm iterates until it does not eliminate any more outliers. We can see the algorithm pseudo-code shown in *Algorithm* 2.

Algorithm 2 Confidence Region algorithm

Require: $I = \{1, \ldots, N\}, \alpha = 10^{-20}$
  while true **do**
    $\hat{y} \leftarrow LinReg\left(\left\{\left(x^{(i)}, y^{(i)}\right)\right\}\right)_{i \in I}$
    $\hat{e}_i \leftarrow y_i - \hat{y}_i \quad \forall i \in I$
    if $y_i < \hat{y}_i$ **then**
      $\theta_i = 2 \cdot \Phi(\frac{(y_i - \hat{y}_i)}{s_i})$ ▷ s_i is the deviation standard of $\hat{y}$
    else
      $\theta_i = 1 - 2 \cdot \Phi(\frac{(y_i - \hat{y}_i)}{s_i})$
    end if
    $J \leftarrow \{i \in \{1...N\} \mid \theta_i \leq \alpha\}$
    $K \leftarrow K \setminus J$
    if $|J| = 0$ **then**
      break
    end if
  end while

3.2.3. Comparison of the outliers detection methods

In the previous section, we presented two methods to address the elimination of outliers. The methods are referred to as the *Interquartile* method and the *Confidence Region* method. We apply both methods in the Hungarian Grand Prix 2022, considering all drivers' lap times (i.e. as if we were at the end of the race).

Figure 3.4 shows the lap times for every lap of the race (as in Figure 3.1), indicating in red the lap times that are labeled as outliers by the Interquartile method (left panel) or the Confidence Region method (right panel). We observed key differences between the two approaches. As we can see, we have some data points that significantly deviate from the rest of the data. These lap times are observations that fall far outside the typical range of values. But, not all of them are removed, only those we can not explain. In the *Interquartile* method, the amount of data points that are eliminated is significantly less than in the *Confidence Region* method.

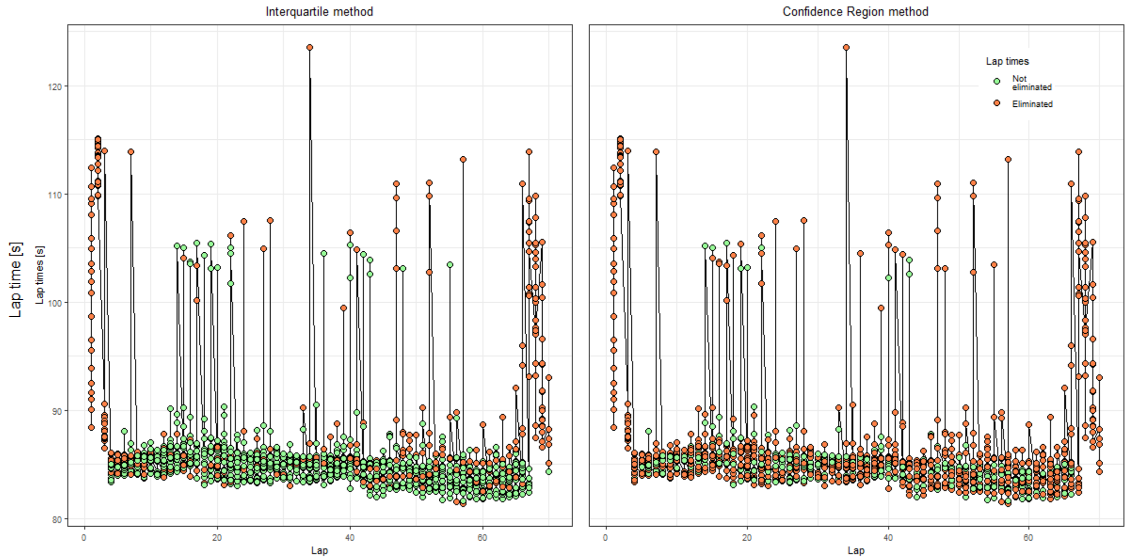

Figure 3.4: Comparison of eliminated lap times for the Hungarian Grand Prix 2022.

Additionally, we graph the residuals of the regression after applying both methods, see Figure 3.5. We see big differences between the observed lap times and the predicted ones on some of the values obtained from the regression model when we apply the *Interquartile* method. As it was said before, this is because it eliminates less data. Note that, unlike the case when we consider all data points (as in Figure 3.3), in this case, the residuals of the points left by the outlier detection methods are considerably smaller in magnitude, especially when using the Confidence Region method (right panel of Figure 3.5)

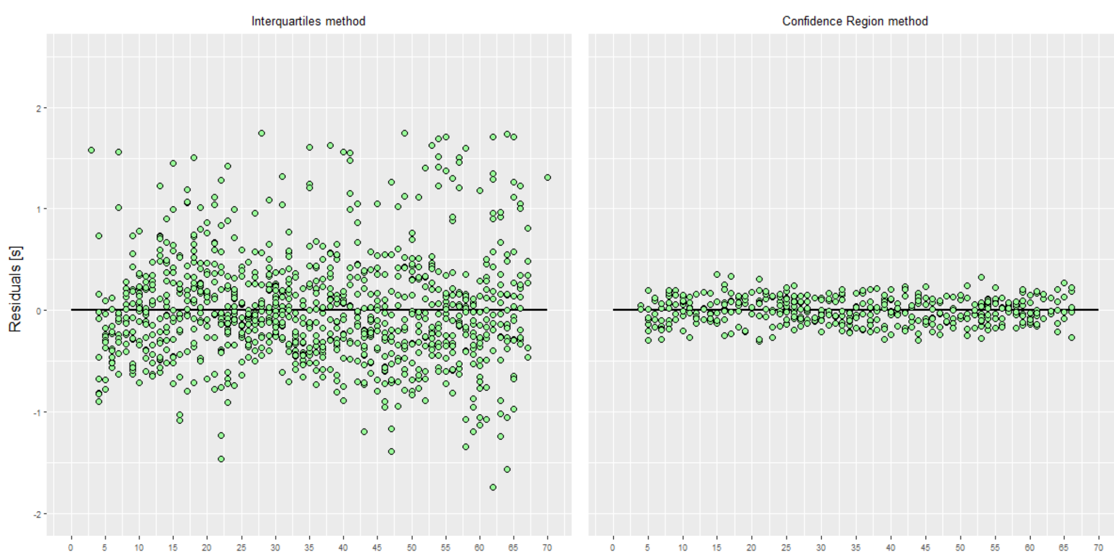

Figure 3.5: Comparison of residuals for both methods for the Hungarian Grand Prix 2022.

Finally, we graph the number of lap times each method eliminates in each iteration. Figure

3.6 shows the number of lap times eliminated (y-axis) in each iteration (x-axis) by each of the outlier detection methods. As seen in the previous figures, the amount of lap times the Confidence Region method eliminates is larger than the *Interquartile* method.

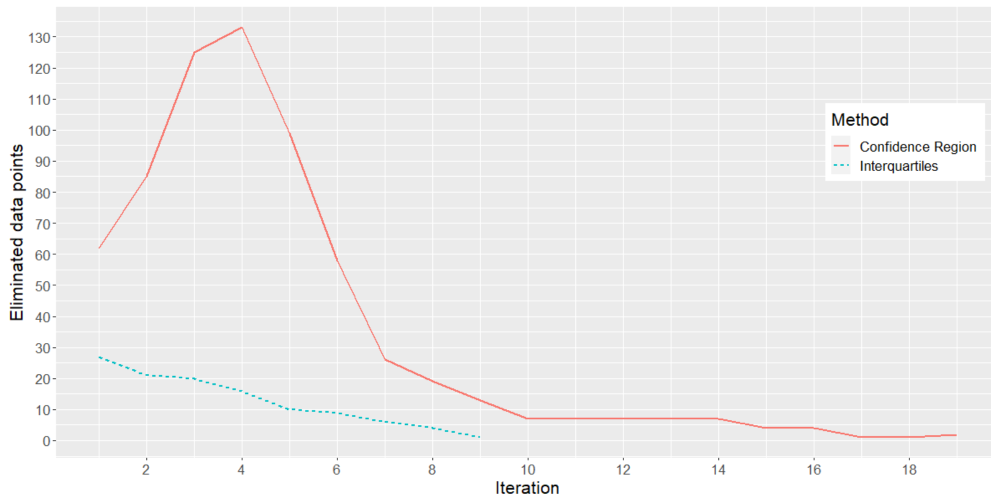

Figure 3.6: Comparison of eliminated data for each method for the Hungarian Grand Prix 2022.

Overall, these visualization and analysis approaches have provided us with valuable information about both methods. Now that we know how they work, we want to compare the results. First, let's take a look at the results obtained from the regression. Table 3.3 shows the results of the regression considering all laps with the *Interquartile* method.

Table 3.3: Summary results of the regression considering all laps with the *Interquartile* method

term	estimate	std.error	statistic	p.value
(Intercept)	83.150	0.126	660.528	< 2e-16
lap	-0.019	0.005	-3.773	< 2e-16
lap^2	0.001	0.000	-3.700	< 2e-16
inpits	2.826	0.094	29.910	< 2e-16
outpits	20.017	0.126	159.177	< 2e-16
team Alfa Romeo	0.831	0.082	10.125	< 2e-16
team Alpha Tauri	1.317	0.077	17.177	< 2e-16
team Alpine	0.824	0.084	9.855	< 2e-16
team Aston Martin	0.777	0.082	9.509	< 2e-16
team Hass	1.290	0.075	17.278	< 2e-16
team McLaren	0.813	0.074	11.058	< 2e-16
team Mercedes	0.076	0.072	1.059	0.290
team Red Bull	-0.135	0.073	-1.858	0.063
team Williams	1.621	0.079	20.398	< 2e-16
time difference non lapped car behind	0.438	0.091	4.799	< 2e-16
time difference non lapped car in front	0.851	0.151	5.641	< 2e-16
time difference lapped car in front	0.162	0.166	0.974	0.330
Medium	0.544	0.099	5.483	< 2e-16
Hard	1.511	0.116	13.016	< 2e-16
Soft·lapwear	0.146	0.008	17.298	< 2e-16
Medium·lapwear	0.061	0.003	17.760	< 2e-16
Hard·lapwear	0.031	0.004	8.227	< 2e-16

The coefficient of determination R^2 obtained is 0.97, with a total of 975 data used.

Based on the p-values, we observe that almost all of the variables are statistically significant predictors of lap time. Mercedes and Red Bull teams, along with the time difference with the lapped car in front, on the other hand, do not show a statistically significant impact on race time.
If we take a look at the team performances, we get that the fastest team is Red Bull, followed by Ferrari, and Mercedes, with a difference of 0.076 seconds by lap between them. We also can see that the slower teams on the grid are Williams and Haas, with more than a 1-second gap to the fastest teams. Thus, we see that the team has a big effect on the lap time.
Finally, we can obtain information about the tire's performance. We have that the Soft tire is the fastest, then we have the Medium, 0.544 seconds slower per lap, and finally the Hard, 1.511 seconds slower per lap, which is consistent with reality. Also, when looking at the slopes, we see that the slopes of all compound tires are positive, with the Hard one being flatter. This is also consistent with reality since the Soft tire has a much shorter useful life, that is, although it is faster, it also wears out faster. Contrary to the medium, which has a lower performance at the beginning but lasts longer. And the Hard, which even though it's

slower at the beginning, lasts longer.

Table 3.4: Summary results of the regression considering all laps with the *Confidence Region* method

term	estimate	std.error	statistic	p.value
(Intercept)	82.953	0.045	1839	< 2e-16
lap	-0.022	0.002	-11.576	< 2e-16
lap^2	0.001	0.000	-8.219	< 2e-16
inpits	2.539	0.031	81.135	< 2e-16
outpits	19.939	0.047	422.798	< 2e-16
team Alfa Romeo	1.149	0.030	37.936	< 2e-16
team Alpha Tauri	1.378	0.028	48.356	< 2e-16
team Alpine	0.856	0.031	27.449	< 2e-16
team Aston Martin	0.578	0.031	18.747	< 2e-16
team Hass	1.264	0.028	44.523	< 2e-16
team McLaren	1.015	0.027	37.446	< 2e-16
team Mercedes	0.115	0.026	4.361	< 2e-16
team Red Bull	-0.136	0.027	-5.082	< 2e-16
team Williams	1.589	0.030	52.221	< 2e-16
time difference non lapped car behind	0.386	0.038	10.217	< 2e-16
time difference non lapped car in front	1.138	0.056	20.216	< 2e-16
time difference lapped car in front	0.140	0.060	2.326	0.020
Medium	0.596	0.038	15,705	< 2e-16
Hard	1.382	0.045	30.607	< 2e-16
Soft·lapwear	0.169	0.038	52.180	< 2e-16
Medium·lapwear	0.061	0.003	47.456	< 2e-16
Hard·lapwear	0.033	0.001	26.206	< 2e-16

The coefficient of determination R^2 obtained is 0.998, with a total of 444 data used.

For the *Confidence Region* method, the results are shown in Table 3.4. Based on the p-values, we observe that all of the variables are statistically significant predictors of lap time. The only one that does not show a statistically significant impact on race time it's the time difference lapped with the car in front.
If we take a look at the team and tire performances, we get almost the same results as in the *Interquartile* method. The key difference between both methods is the deviation standard of the tire performance. This method has almost half of the deviation standard the other method gets for the compound tires.

Continuing with the analysis, from the coefficients obtained in the regression for each method, taking a new Soft tire as a reference, we plot the performance of all tire compounds, obtaining the additional time per lap for different values of tire wear. Figure 3.7 shows the estimated tire compound performance (y-axis), in terms of additional lap time in comparison to a new Soft compound, for different levels of wear (x-axis). The left panel of Figure 3.7 shows the case when removing outliers with the *Interquartile* method, and the right panel shows the case when removing outliers with the *Confidence Region* method.

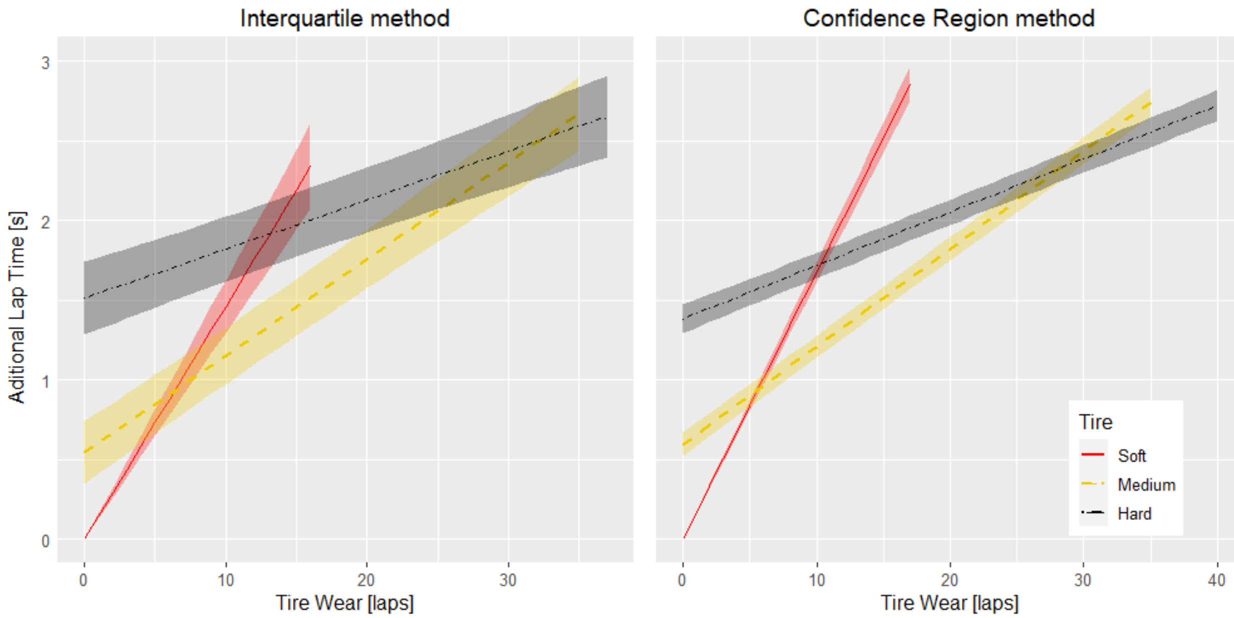

Figure 3.7: Comparison of additional time per lap with each tire for the Hungarian Grand Prix 2022.

As we can see in Figure 3.7, one of the key differences between both methods is the deviation standard of the tire performance. Because the *Confidence Region* method uses fewer data, it has little deviation in comparison. The standard deviation is a measure of spread that indicates how far individual values are from the mean of the data set. When the standard deviation is smaller, it means that the values are more clustered. The standard deviation obtained is due to the values being closer to the mean. Outliers can increase the standard deviation as they move away from the mean. If there are few or no outliers in the data set, the standard deviation will be smaller. As mentioned above, since this method removes more data compared to the other method, its standard deviation is smaller.
About the results themselves, we can see that they are very similar, the Soft compound tire is faster when it has less wear, but it increases its lap time as the use increases. For the Hard tire, we have that is slower at the beginning, but its performance tends to decrease at a lower rate compared to the other tire compounds.

So far, we have estimated the tire performance using all the lap times of a race, except lap times labeled as outliers. However, in reality, if we were to estimate the tire performance in the middle of the race, i.e., in lap 15, we would not be able to use data from future laps since those lap times are still unknown. Figure 3.8 shows the estimated tire compound performance (y-axis) at lap 15 of the race, in terms of additional lap time in comparison to a new Soft compound, for different levels of tire wear (x-axis). The left panel of Figure 3.8 shows the Interquartile method and the right panel shows the Confidence Region method.

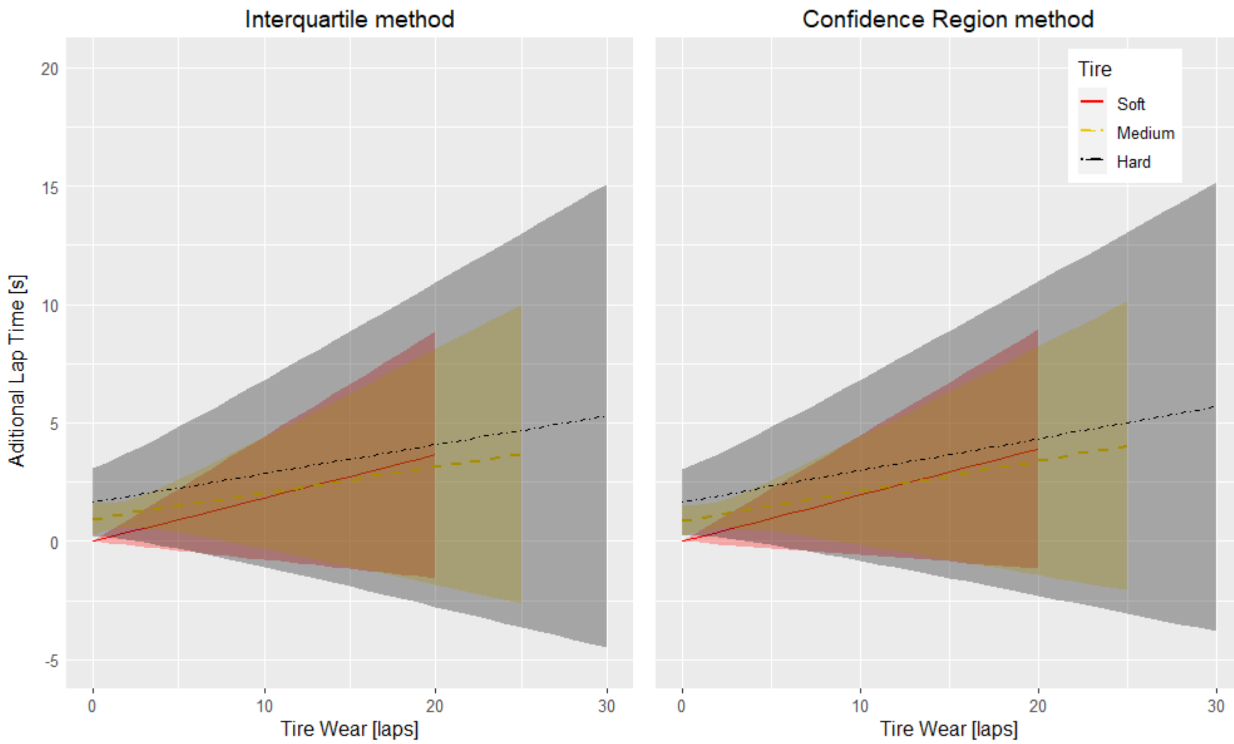

Figure 3.8: Comparison of additional time per lap with each tire at lap 15 for the Hungarian Grand Prix 2022.

As we can see, for both methods the results are similar but not very accurate. We have a very high standard error for all tire compounds' performance, showing a large amount of variation in the group that is being studied. This is because the cars are just adjusting to the track conditions and the drivers are getting used to the performance of the tires, so there is no clear trend to know the behavior of the tires, nor their performance. Thus, it is very difficult to make reliable predictions on the first stages of the race, because there is not much data available.

Based on the given information, we can see that the results of both methods are quite similar. However, there are two key differences between them. Firstly, the *Interquartile* method yields higher standard errors compared to the *Confidence Region* method. Secondly, the *Confidence Region* method eliminates a larger number of data points, particularly lap times, than the *Interquartile* method. Thus, since the dataset we are working with has a limited amount of data, we want to work with a method that delivers reliable results but without eliminating a large amount of data, is that we chose to continue working with the *Interquartile* method. It is also important to consider that this method is a robust technique for identifying outliers because it is not sensitive to extreme values. It focuses on the spread of data within the middle 50% (between the first quartile and the third quartile) and is less affected by extreme values or outliers. And because some potential outliers are genuine data points, the *Interquartile* method is the better choice.

3.3. Coefficient priors

Up to this point, we have been working with data from the entire race. But, as said before, teams only have data up to the moment they evaluate the strategy. The problem is that in the early stages of the race, there is too little data to make reliable predictions. To improve this, data from the previous year's race at the same circuit can be used to improve the predictions.

In Bayesian statistics, priors refer to initial information about the parameters of a statistical model before any data is observed. In this case, even before the race begins, we have information about the previous year's race. Priors play a fundamental role in Bayesian inference, as they represent the incorporation of existing knowledge into the analysis. These prior beliefs are then combined with the observed data to update our knowledge about the parameters, resulting in a posterior distribution.

Bayesian statistics provides a powerful framework for updating our beliefs based on observed data. The posterior distribution, $p(\theta \mid \text{data})$, is obtained by combining the prior distribution with the likelihood function, which represents the probability of observing the data given the parameters. Bayes' theorem states that the posterior distribution (that is, the up-to-date distribution) is proportional to the product of the "prior" distribution and the likelihood function (which describes how well the data fits the hypothesis or model).
The general formula for Bayes' theorem is as follows:

$$p(\theta \mid \text{data}) = \frac{p(\text{data} \mid \theta) \cdot p(\theta)}{p(\text{data})}$$

where:

- $p(\theta \mid \text{data})$ is the posterior distribution, representing our updated beliefs about the parameters after observing the data available from the race.
- $p(\text{data} \mid \theta)$ is the likelihood function, representing the probability of observing the data given the parameters.
- $p(\theta)$ is the prior distribution, representing the coefficients obtained from the previous year's race. This allows us to have more knowledge about the parameters in the initial stages, and to make better predictions.
- p (data) is the marginal likelihood or evidence, which serves as a normalization constant and ensures that the posterior distribution integrates to 1.

Is important to note that as more data is incorporated, the "prior" tend to have less influence on the posterior distribution, and the estimation is increasingly based on observed data. The prior will have a bigger influence mostly on the early stages of the race, when there is not much data available. In other words, as the race progresses, the data from the actual race will have more influence on the results than the prior.

We decided not to use the data from the previous race simultaneously in a regression with the data from the current race because they are not comparable. Many factors change from one race year to another, such as different specifications for the cars, different climates, and different used tires, among others. Instead, we prefer to use the information from the previous race to generate a prior distribution of the coefficients, which is updated with the partial data from the current race.

It is important to mention that when we evaluate the 2022 Hungarian Grand Prix, we will use the 2020 race as prior. We cannot use the race from 2021 because it rained during that race. Instead, we will use the 2020 race. And even though it rained during that race as well, since the teams used rain tires for the first 4 laps, and considering that the effect of rain lasts for more, the first 8 laps of that race were eliminated.

Consider the case in which we would like to predict the tire performance at lap 15 (of the Hungarian GP 2022). Figure 3.9 shows the data available (y-axis), for different levels of wear (x-axis) using the *Interquartile* method. The left panel of Figure 3.9 shows the case when we only have data for the first 15 laps of the race and the right panel shows the data available of the previous year's race. As we can see in Figure 3.9, we do not have much information about the performance of the tires for the first 15 laps of the race. For example, if we look at the Hard tire, there are not many cars using it. But, if look at the race from the year before, we have a lot more data available, for all tires. In short, having more data available is beneficial because it increases the reliability of the results, reduces error, and improves the precision of the estimates. Having data from the previous year in Formula 1 is essential for effective strategy planning. This data allows teams to learn from the previous season and continue to improve their performance. It is important to consider that when we choose the prior, this must be the race of the previous year, but on the same circuit as the race we are currently working with. The reason why we do not use data from other circuits is that races are very heterogeneous concerning lap times, tire degradation, and specific tire compounds, among others.

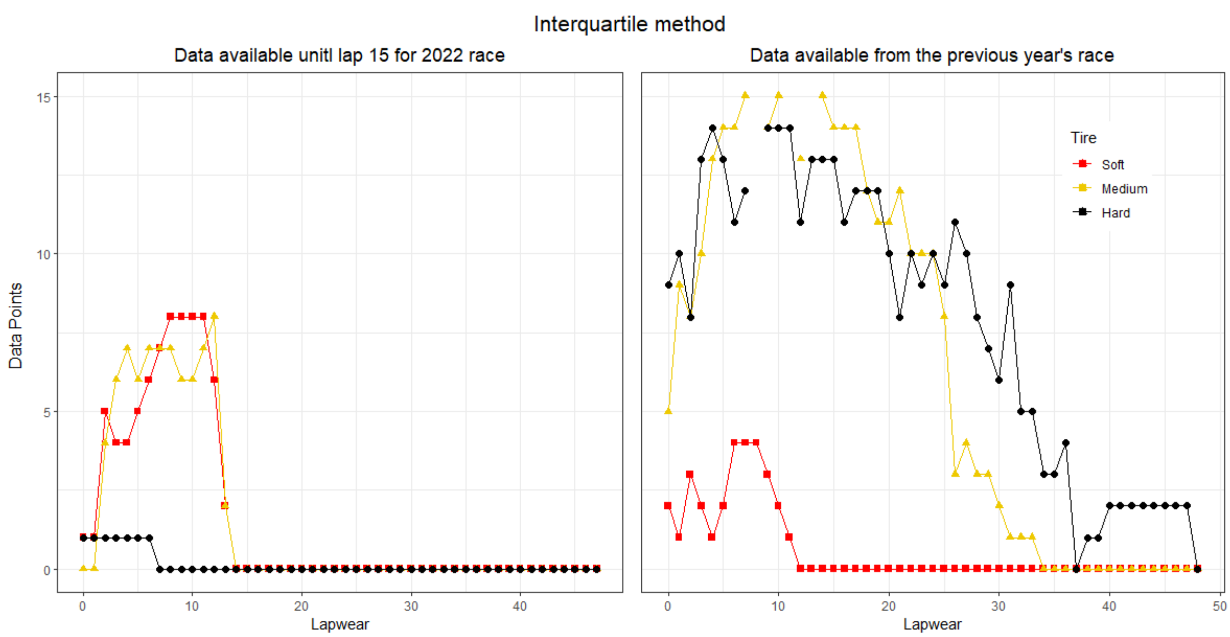

Figure 3.9: Data available for the Hungarian Grand Prix 2022 with the Interquartile method.

To apply the methodology explained above, the regression must first be applied to the race of the previous year. The coefficients and deviations of said regression will be used as prior for the calculation of the regression of the current year. We can see the algorithm pseudo-code shown in *Algorithm* 3.

Algorithm 3 Coefficient prior algorithm

Require: $I = \{1, \ldots, N\}$, current lap n, lap times up to current lap $\left(x^{(i)}, y^{(i)}\right)$, all lap times of previous race $\left(x^{i}_{(p)}, y^{i}_{(p)}\right)$

$\beta^{(prior)}, \sigma_{\beta}^{(prior)} \leftarrow LinReg\left(\left\{\left(x^{(i)}, y^{(i)}\right)\right\}_{i \in I}\right)$

$\beta \leftarrow PosteriorLinReg(\left(\left\{\left(x^{(i)}, y^{(i)}\right)\right\}_{i \in I}, \beta^{(prior)}, \sigma_{\beta}^{(prior)}\right)$

To apply the algorithm, the *stan_glm* function that performs full Bayesian estimation was used in the *R* code. This function allows us to compute generalized linear modeling with optional prior distributions for the coefficients, intercept, and auxiliary parameters.

The model will now be used to improve the results previously obtained. In Figure 3.10 we can see the difference between considering the data from the previous year in the coefficient prior and not. On the left, we have the tire performance for each compound by considering only data from the first 15 laps of the race. In the image in the center, we can see the tires' performance calculated from the previous year's race data. And, in the image to the right we see again the analysis of the performance of the tires on lap 15 of the race, but this time, considering the previous year's race as prior. When we make predictions about tire performance by only considering data from the first 15 laps of the race, the results are not satisfactory, we have a high standard deviation for all the tires, showing a large amount of variation. That's why we use the Hungarian Grand Prix from 2020 as prior. As we have more data available, the results will improve.

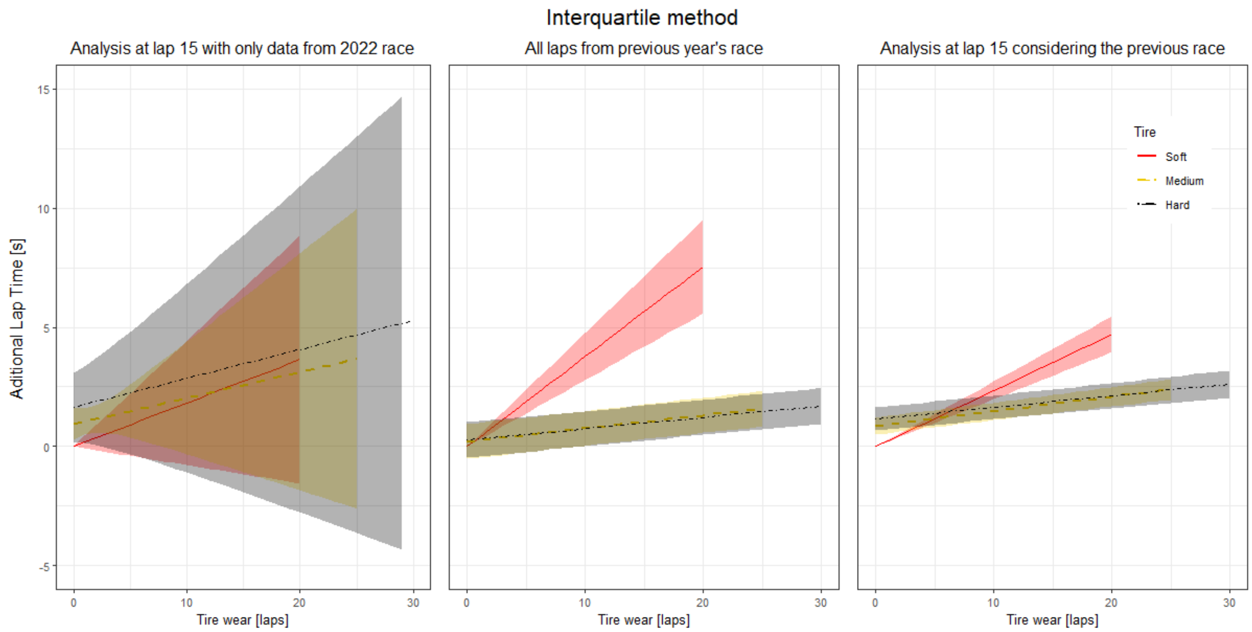

Figure 3.10: Comparison of additional time per lap with each tire for the Hungarian Grand Prix 2022 Interquartile method.

As we can see in the Figure above, the race from the previous year has a big impact on the prediction of tire performance. This is because of the amount of data we have available on the first 15 laps of the race. When using the race from the previous year as prior, we get a massive improvement in the predictions about the tire performance for all compounds during the race. This Bayesian model allows us to increase the amount of data we are working with, and to improve the quality of the results. Now we can analyze tire performance and strategy, not only in the final stages of the race. Since we have more data available, we can also predict tire performance in the early stages of the race.

Chapter 4

Optimization of race strategy

As it was mentioned above, the race strategy may vary depending on factors such as the performance of the car, the characteristics of the circuit, and the specific tactical decisions of each team. Making the right decisions and having the ability to adapt to circumstances in real time are key to success in Formula 1. Sudden changes in track conditions require teams and drivers to be quick to adapt and make decisions. The ability to read the track and understand how the conditions affect the car's performance is crucial. Teams must maintain constant communication with the driver to provide updated information on conditions and adjust strategy accordingly. The race strategy involves making strategic decisions regarding tire usage, pit stops, fuel management, and overall race pace. In this chapter, we will just focus on the tires, by evaluating the best tire strategy for the race, considering the information available at the moment, and updating it as the race progresses. The tire strategy involves deciding how many pits stops to do, when to do them, and what type of tire compound would be used during the different stints.

To start with the optimization, we fix the lap in which we are, n, and we run the regression with the methodology explained in Chapter 3, obtaining the Beta coefficients. We are interested in the best strategy for the remainder of the race. For that, we can list all the possible solutions, using between 0 and k stops. For this, we need to know how long the race time of a strategy profile is(from the current lap to the end of the race). Note that the only thing that differentiates the different strategy profiles, regarding the remaining race time, is: (i) Pit stop time, (ii) Additional time due to tire wear (compared to a new Soft), (iii) Additional time due to tire compound. Once we have the time for each pit stop strategy profile, we can optimize by looking for the best strategy (for the remainder of the race). Finally, we can analyze what strategy would give the methodology for a race considering different laps. In summary,

- Section 4.1: computes the expected time in the remainder of the race for a fixed strategy profile.
- Section 4.2: shows the optimization of the best running profile in what remains.
- Section 4.3: shows numerical results of the evaluation of the method for different laps of the race.

4.1. Estimate race time of a profile strategy

To optimize the best race strategy, we want to minimize the total race time for the driver we are analyzing. The objective is that the driver completes the race distance in the shortest possible time by using the best tire strategy. Since we want to compute the expected time in the remainder of the race for a fixed strategy profile, the following variable will be defined, Z_{tk}. The additional time per lap when using a t tire compound, with k laps of usage, compared to a new Soft tire compound as a reference.

$$Z_{tk} = \beta_{t0} + \beta_{t1} k$$

Where, $t \in \mathcal{T} = \{\text{Soft}, \text{Medium}, \text{Hard}\}$ is the set of possible tires, $k \in \mathcal{N}$, set of laps of the race, and the Betas are random variables. Note that since this variable is constructed from the β obtained from the regression, is expected to follow a normal distribution. Therefore, $Z_{tk} \sim N(\mu_{Z_{tk}}, \sigma^2_{Z_{tk}})$. First, for the mean we have,

$$\mu_{Z_{tk}} = \hat{\beta}_{t0} + \hat{\beta}_{t1} k$$

Where, $\hat{\beta}_{t0}$ and $\hat{\beta}_{t1}$ are the coefficients obtained from the regression. And, for the variance:

$$\sigma^2_{Z_{tk}} = Var\left(\beta_{t0} + \beta_{t1} k\right) \tag{4.1}$$

$$= Var\left(\beta_{t0}\right) + 2 \cdot k \cdot Cov\left(\beta_{t0}, \beta_{t1}\right) + k^2 \cdot Var\left(\beta_{t1}\right) \tag{4.2}$$

Following that, we can express $\mu_{Z_{tk}}$ and $\sigma^2_{Z_{tk}}$ using matrix notation. To achieve this, we define:

$$a_k = \begin{pmatrix} 1 \\ k \end{pmatrix}, b_t = \begin{pmatrix} \beta_{t0} \\ \beta_{t1} \end{pmatrix}, \Sigma^{(t)} = \begin{bmatrix} Var\left(\beta_{t0}\right) & Cov\left(\beta_{t0}, \beta_{t1}\right) \\ Cov\left(\beta_{t1}, \beta_{t0}\right) & Var\left(\beta_{t1}\right) \end{bmatrix}$$

With $\Sigma^{(t)}$ the covariance matrix of the coefficients of tire compound t. Consequently, we obtain:

$$\mu_{Z_{tk}} = a_k^T b_t$$

$$Var\left(Z_{tk}\right) = a_k^T \Sigma^{(t)} a_k$$

The variable we just defined, Z_{tk}, allows us to get the additional lap time with each tire. But, since we want to evaluate a complete stint, we need to sum all the Z_{tk} that belong to this stint. A stint refers to the period a driver spends on the track continuously without pitting to change tires. Making the right decisions regarding stints can significantly impact a driver's position and performance throughout the race.

Consider a single stint in which the driver is using the tire compound t; so that at the beginning of the stint the tire has a wear of i laps, whereas by the end of the stint, the tire wear equals $f > i$. Thus, the stint has a duration of $f - i + 1$ laps. Consequently, we define a new variable, W_{tif}, that will consider all laps in the stint we are evaluating, by giving the additional race time during the stint(compared to using a new Soft compound in each of these laps).

$$W_{tif} = \sum_{k=i}^{f} Z_{tk}$$

A strategy profile consists of the tire strategy for a specific amount of stops. Thus, we need to describe the stints by considering tire compound, tire wear, and usage. To describe them, we will work with the following variables. First of all, we define $t \in \mathcal{T} = \{\text{Soft}, \text{Medium}, \text{Hard}\}$. This variable corresponds to the tire compound being used by the driver. Also, we define w, a variable that provides the number of laps that the tire already has when it starts being used. Also, we got u, the number of laps that the selected tire will be used in the stint. Finally, as it was defined before, we have the set $\mathcal{N} = \{1, \dots, N\}$, the number of laps of the race.

With the variables we just created, we can express W_{tif} as W_{twu}. Where $i = w$, the tire wear at the beginning of the stint, $f = w + u - 1$, the duration of the stint. Thus,

$$W_{twu} = \sum_{k=w}^{w+u-1} Z_{tk} \tag{4.3}$$

$$= \left(\sum_{k=w}^{w+u-1} a_k \right)^T b_t \tag{4.4}$$

$$= \begin{pmatrix} u \\ \frac{(w+u-1)(w+u)}{2} - \frac{w(w-1)}{2} \end{pmatrix}^T b_t \tag{4.5}$$

We now have a variable, W_{twu}, that describes the additional time for a stint, compared to a new Soft tire. Note that the variables Z_{tk} are not independent. Therefore, to evaluate the distribution of W_{twu}, a multivariate normal distribution must be considered, allowing us to work with a set of correlated random variables. Since we have that $W_{twu} = (\sum_{k=w}^{w+u-1} a_k)^T b_t$, and we know that the coefficient estimators β obtained through the regression can approximate a normal distribution, we have that $W_{twu} \sim N(\mu_{W_{twu}}, \sigma^2_{W_{twu}})$. Consequently, W_{twu} follows a multivariate normal distribution. Where,

$$\mu_{W_{twu}} = \mathbb{E}\left[W_{t_s w_s u_s}\right] \tag{4.6}$$

$$= \left(\sum_{k=w}^{w+u-1} a_k \right)^T \mathbb{E}\left[b_t\right] \tag{4.7}$$

$$= \left(\sum_{k=w}^{w+u-1} a_k \right)^T \begin{pmatrix} \hat{\beta}_{t_0} \\ \hat{\beta}_{t_1} \end{pmatrix} \tag{4.8}$$

And,

$$\sigma^2_{W_{twu}} = Var\left(W_{twu}\right) \tag{4.9}$$

$$= Var\left(\sum_{k=w}^{w+u-1} a_k^T b_t \right) \tag{4.10}$$

We are now able to calculate the mean and variance for a specific stint within a race. However, as our objective is to optimize the race strategy, we must sum all stints throughout the race.

Let's take Max Verstappen's strategy during the 2022 Hungarian Grand Prix as an example. He started his race with a 2-lap used soft tire and used it until lap 16, where he made a pit stop, and put on a fresh medium tire. This stint lasted until lap 38, where he changed to another new medium tyre, and used it until the end of the race.

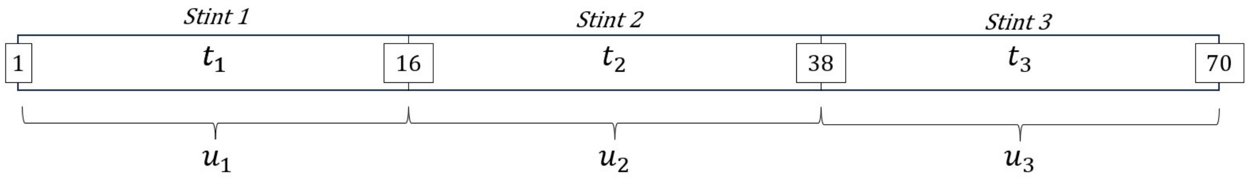

Figure 4.1: Max Verstappen's stints during Hungarian Grand Prix 2022.

By defining $\mathcal{S} = \{1, \ldots, S\}$ as the number of stints during the race, we can now sum all the stints a driver makes during the race. Before computing the sum, let us define $c_t \in \mathbb{R}^2$,

$$c_t = \sum_{s=1}^{\mathcal{S}} \left[1_{\{t_s = t\}} \sum_{k=w_s}^{w_s+\mu_s-1} a_k \right] \qquad \forall t \in \mathcal{T}$$

And,

$$c = \begin{pmatrix} c_{soft} \\ c_{medium} \\ c_{hard} \end{pmatrix} \in \mathbb{R}^6,\ b = \begin{pmatrix} b_{soft} \\ b_{medium} \\ b_{hard} \end{pmatrix} \in \mathbb{R}^6$$

In addition, let us denote by Σ_b the variance-covariance matrix of b, namely,

$$\Sigma_b = \begin{bmatrix} Var(\beta_{s0}) & Cov(\beta_{s0}, \beta_{s1}) & Cov(\beta_{s0}, \beta_{m0}) & Cov(\beta_{s0}, \beta_{m1}) & Cov(\beta_{s0}, \beta_{h0}) & Cov(\beta_{s0}, \beta_{h1}) \\ Cov(\beta_{s1}, \beta_{s0}) & Var(\beta_{s1}) & Cov(\beta_{s1}, \beta_{m0}) & Cov(\beta_{s1}, \beta_{m1}) & Cov(\beta_{s1}, \beta_{h0}) & Cov(\beta_{s1}, \beta_{h1}) \\ Cov(\beta_{m0}, \beta_{s0}) & Cov(\beta_{m0}, \beta_{s1}) & Var(\beta_{m0}) & Cov(\beta_{m0}, \beta_{m1}) & Cov(\beta_{m0}, \beta_{h0}) & Cov(\beta_{m0}, \beta_{h1}) \\ Cov(\beta_{m1}, \beta_{s0}) & Cov(\beta_{m1}, \beta_{s1}) & Cov(\beta_{m1}, \beta_{m0}) & Var(\beta_{m1}) & Cov(\beta_{m1}, \beta_{h0}) & Cov(\beta_{m1}, \beta_{h1}) \\ Cov(\beta_{h0}, \beta_{s0}) & Cov(\beta_{h0}, \beta_{s1}) & Cov(\beta_{h0}, \beta_{m0}) & Cov(\beta_{h0}, \beta_{m1}) & Var(\beta_{h0}) & Cov(\beta_{h0}, \beta_{h1}) \\ Cov(\beta_{h1}, \beta_{s0}) & Cov(\beta_{h1}, \beta_{s1}) & Cov(\beta_{h1}, \beta_{m0}) & Cov(\beta_{h1}, \beta_{m1}) & Cov(\beta_{h1}, \beta_{h0}) & Var(\beta_{h1}) \end{bmatrix}$$

Thus, we get the following:

$$\mathbb{E}\left[\sum_{s=1}^{\mathcal{S}} W_{t_s w_s u_s}\right] = \sum_{s=1}^{\mathcal{S}} \left(\sum_{k=w_s}^{w_s+u_s-1} a_k \right)^T \mathbb{E}\left[b_{t_s}\right] \tag{4.11}$$

$$= \sum_{s=1}^{\mathcal{S}} \left(\sum_{k=w_s}^{w_s+u_s-1} a_k \right)^T \begin{pmatrix} \hat{\beta}_{t_s 0} \\ \hat{\beta}_{t_s 1} \end{pmatrix} \tag{4.12}$$

$$= c^T b \tag{4.13}$$

Then, the variance can be expressed as

$$\begin{aligned} Var\left(\sum_{s=1}^{\mathcal{S}} W_{t_s w_s u_s}\right) &= Var\left(\sum_{s=1}^{\mathcal{S}} \sum_{k=w_s}^{w_s+u_s-1} a_k^T b_{t_s}\right) && (4.14)\\ &= Var\left(\sum_{t\in\mathcal{T}} c_t^T b_t\right) && (4.15)\\ &= Var\left(c^T b\right) && (4.16)\\ &= c^T \Sigma_b c && (4.17)\end{aligned}$$

As a result, we can now estimate the race time for a specific strategy.

4.2. Optimization of the best running profile

We are interested in the best strategy for the remainder of the race. From the information provided earlier, we now have the capability to determine a strategy profile's race time, from the current lap to the end of the race. Thus, we can start with the optimization, by identifying the strategy that incurs the shortest time for the remainder of the race.

We show that the optimization problem of finding the best pit stop strategy for the remainder of the race can be formulated as a Mixed-Integer Quadratic Programming, MIQP. It is a type of optimization problem that involves finding the values of continuous and integer variables to minimize a quadratic objective function while satisfying a set of constraints. In MIQP, some of the decision variables are restricted to be integers, which makes the problem more complex compared to linear programming.

Consider the case that there are n laps left until the end of the race, and we are currently using tire compound $t_0 \in \mathcal{T}$ which have been used for w_0 laps. For each future tire stint, we would like to know which tire compound will be used, and for how many laps. Before the formulation of the optimization problem, we will provide a non-trivial upper bound on the number of stints for the remainder of the race.

Proposition: The maximum number of pit stops that the optimization problem handle is:

$$k = \left\lceil \frac{n}{\left\lfloor \sqrt{\frac{\beta_{pits}}{\bar{\beta}_1}} \right\rfloor} \right\rceil, \qquad (4.18)$$

where $\bar{\beta}_1 = \max_{t\in\tau} \beta_{t1}$ and $\beta_{pits} = \hat{\beta}_{outpits} + \hat{\beta}_{inpits}$.

Proof: Consider for a moment a hypothetical setting in which there is an infinite number of laps remaining. In addition, consider for a moment there is a single tire compound, with intercept and slope equal to β_0 and β_1, respectively. We would like to see what is the minimum number of laps for a stint, in the sense that having a stint with fewer laps would not be convenient. In particular, we would like to minimize the average time per lap. Let m

be the number of laps of a stint. The accumulated time until lap m can be computed as:

$$\sum_{i=0}^{m-1} \beta_0 + i\beta_1 \quad = \quad m\beta_0 + \beta_1 \frac{(m-1)m}{2} \tag{4.19}$$

Then, the average time per lap considering a pit stop can be computed as:

$$\frac{m\beta_0 + \beta_1(m-1)m + \beta_{pits}}{m} = \beta_0 + (m-1)\beta_1 + \frac{\beta_{pits}}{m} \tag{4.20}$$

Taking the derivative of the RHS of Equation (4.20), we get,

$$\frac{\mathrm{d}}{\mathrm{d}m^\star}\left(\beta_0 + (m^\star - 1)\beta_1 + \frac{\beta_{pits}}{m^\star}\right) = 0 \tag{4.21}$$

$$\beta_1 - \frac{\beta_{pits}}{(m^\star)^2} = 0 \tag{4.22}$$

$$m^\star = \left\lfloor \sqrt{\frac{\beta_{pits}}{\bar{\beta}_1}} \right\rfloor \tag{4.23}$$

Note that we take the floor function since we are looking for an upper bound on the number of stops; thus, a lower bound in the laps of a stint. Then, the maximum number of remaining stints can be computed as the remaining number of laps, divided by Equation (4.23), namely,

$$\left\lceil \frac{n}{m^\star} \right\rceil = \left\lceil \frac{n}{\left\lfloor \sqrt{\frac{\beta_{pits}}{\beta_1}} \right\rfloor} \right\rceil \tag{4.24}$$

If we now consider all possible tire compounds, we look to the worst-case scenario, which corresponds to the compound with the largest β_1. Then, the maximum number of stints can be upper bounded by

$$k = \left\lceil \frac{n}{\left\lfloor \sqrt{\frac{\beta_{pits}}{\bar{\beta}_1}} \right\rfloor} \right\rceil \tag{4.25}$$

where $\bar{\beta}_1 = \max_{t \in \mathcal{T}} \beta_{t1}$ and $\beta_{pits} = \hat{\beta}_{outpits} + \hat{\beta}_{inpits}$.

□

From the Proposition above, k denotes the maximum number of stints. It is also important to consider the parameter γ, which indicates if, during the current lap, there is a yellow flag.

To start with the optimization, we define the following decision variables:

$x_{s,t}$: the number of laps in which the potential tire stint $s \in \{0, 1, \ldots, k\}$ with the tire $t \in \mathcal{T}$ is used.

$y_{s,t}$: 1 if the potential tire stint $s \in \{1, \ldots, k\}$ with the tire $t \in \mathcal{T}$ is used for at least one lap, 0 otherwise.

z : 1 if $x_{0,t} = 0$, and $\gamma = 1$ (i.e., there is a yellow flag), 0 otherwise. This is, if a pitstop is made under a yellow flag.

Note that the potential stints are not guaranteed to occur in a specific strategy profile for the remainder of the race. For example, if there are 40 laps left, and we are currently using a Medium compound, we can stop in ten more laps for a Soft tire compound, and then stop in 30 more laps for a Hard tire compound. Then, this strategy could be represented with the decision variables taking values $v = 10,\ x_{1,Soft} = 20,\ x_{2,Hard} = 10,\ y_{1,Soft} = y_{1,Hard} = 1$, and zero for all the other decision variables. However, the same example can also be represented with other solutions, such as $v = 10,\ x_{3,Soft} = 20,\ x_{4,Hard} = 10,\ y_{3,Soft} = y_{4,Hard} = 1$. This motivates introducing constraints that reduce the symmetry of the problem.

The objective function is the remaining race time relative to the hypothetical case of racing all future laps with new Soft tires in each lap with no pit stops. Consider a fixed potential tire stint s that is raced during x_s laps. Then,

$$
\begin{aligned}
\mu_{W_{t_s w_s x_s}} &= \sum_{k=w_s}^{w_s+x_s-1} a_k^T \begin{pmatrix} \hat{\beta}_{t_s 0} \\ \hat{\beta}_{t_s 1} \end{pmatrix} && (4.26) \\
&= \begin{pmatrix} x_s \\ \frac{(w_s+x_s-1)(w_s+x_s)}{2} - \frac{w_s(w_s-1)}{2} \end{pmatrix}^T \begin{pmatrix} \hat{\beta}_{t_s 0} \\ \hat{\beta}_{t_s 1} \end{pmatrix} && (4.27) \\
&= \begin{pmatrix} x_s \\ \frac{2w_s x_s + x_s^2 - x_s}{2} \end{pmatrix}^T \begin{pmatrix} \hat{\beta}_{t_s 0} \\ \hat{\beta}_{t_s 1} \end{pmatrix} && (4.28) \\
& && (4.29) \\
&= (\hat{\beta}_{t_s 0} + \hat{\beta}_{t_s 1}(w_s - \frac{1}{2}))x_s + \frac{\hat{\beta}_{t_s 1}}{2} x_s^2 && (4.30)
\end{aligned}
$$

Since the goal of the problem is to minimize the time it takes to finish the race, we would work with an objective function that combines linear and quadratic terms based on β_0 and β_1, pit stop time, and decisions made at each pit stop. The goal is to find the pit stop configuration that minimizes this function, thus optimizing the race strategy.

From the $\mu_{W_{t_s w_s x_s}}$ recently calculated, we can build the objective function. In addition to the above, the pit stop time must be added throughout the term, $\beta_{pits} \cdot y_{s,t}$. Where $\beta_{pits} = \hat{\beta}_{outpits} + \hat{\beta}_{inpits}$. Also, we would consider the scenario in which during the current lap there is a yellow flag, by subtracting half the time it takes to make a pit stop.
Thus, we get the following MIQP:

$$
\min \sum_{s=0}^{k} \sum_{t \in \mathcal{T}} (\hat{\beta}_{t0} + \hat{\beta}_{t1}(w_t - \frac{1}{2}))x_{s,t} + \frac{\hat{\beta}_{\tau_t 1}}{2} x_{s,t}^2 + \sum_{s=1}^{k} \sum_{t \in \mathcal{T}} y_{s,t} \cdot \beta_{pits} - \frac{\beta_{pits}}{2} z
$$

$$
\sum_{s=0}^{k} \sum_{t \in \mathcal{T}} x_{s,t} = n \tag{4.31}
$$

$$
x_{s,t} - n \cdot y_{s,t} \leq 0 \ \ \forall t \in \mathcal{T},\ \forall s \in \{0, 1, \ldots, k\} \tag{4.32}
$$

$$x_{s,t} - y_{s,t} \geq 0 \ \ \forall t \in \mathcal{T}, \ \forall s \in \{0, 1, \dots, k\} \tag{4.33}$$

$$\sum_{t \in \mathcal{T}} y_{s,t} \geq 1 \quad \text{if} \, |t \cup t_0| = 1, \ \ \forall s \in \{1, \dots, k\} \tag{4.34}$$

$$\sum_{t \in \mathcal{T}} y_{1,t} \leq 1 \tag{4.35}$$

$$\sum_{t \in \mathcal{T}} x_{0,t} = v \tag{4.36}$$

$$\sum_{t \in \mathcal{T}} y_{s,t} \geq \sum_{t \in \mathcal{T}} y_{s+1,t} \quad \forall s \in (1, ..., k-1) \tag{4.37}$$

$$z \leq \left(1 - \frac{v}{n}\right) \gamma \tag{4.38}$$

$$\gamma \, (1 - v) \leq z \tag{4.39}$$

$$v \geq 1_{\{n=N\}} \tag{4.40}$$

$$x_{s,t} \geq x_{s+1,t} \quad \forall t \in \mathcal{T}, \forall s \in \{1, \dots, k-1\} \tag{4.41}$$

$$n \, \gamma + v + w_0 \geq x_{1,t} \cdot 1_{\{t_0 = \text{ Soft}\}} + x_{1,t} \cdot 1_{\{t_0 = \text{ Medium}\}} + x_{1,t} \cdot 1_{\{t_0 = \text{ Hard}\}} \tag{4.42}$$

$$x_{s,t} \geq 0 \quad \forall t \in \mathcal{T}, \ \forall s \in \{0, 1, \dots, k\} \tag{4.43}$$

$$y_{s,t} \in \{0, 1\} \quad \forall t \in \mathcal{T}, \ \forall s \in \{1, \dots, k\} \tag{4.44}$$

$$z \in \{0, 1\} \tag{4.45}$$

In the context of the optimization problem just defined, constraints play a crucial role in defining the boundaries and limitations within which the solution must operate.

- In Equation 4.31 we have that the sum of the number of laps in which the potential tire stint is used must equal the total number of remaining laps n.
- Equation 4.32 shows that if there is a pit stop, the tire can have a usage greater than 0. If there is no pit stop, the tire that would correspond to that stop cannot have usage.
- Equation 4.33 represents that if a tire change occurs, the corresponding tire can have usage.
- Equation 4.34 shows that if only one tire type has been used previously, a tire change must occur in at least one pit stop.

As it was mentioned before, we need constraints that reduce the symmetry of the problem. Thus, we have Equation 4.35 and Equation 4.37. That shows that only if the tuple representing the pitstop before had been used, there can be a new pitstop. This is, the tuples must be used in an ordered manner.

Then, we would be considering the scenario where there is a yellow flag during the lap we are evaluating the strategy. This is, if the driver decides to make a pit stop during the current lap, under a yellow flag, there would be subtracted from the objective function half the pit stop time. During a yellow flag period, all cars on the track are required to slow down. This slower pace allows teams to execute pit stops more comfortably and safely. Overall, the combination of reduced speeds, no overtaking, predictable timing, an extended pit window, and a focus on safety rather than competition contributes to faster pit stops under yellow flag conditions. This is why we reduce the time of a pitstop in half. To introduce this in the MIQP, we need constraints that make relations between γ, v and z, that's what we can see in constraints 4.38 and 4.39.

Finally,

- Equation 4.40 shows that during the first lap of the race, there can not be any change in the tires, this is, if a team chooses a tire to start the race with, it must be used for at least one lap.
- Equation 4.41 is to ensure that if we have 2 stints with the same tire, the first one would be longer than the second one. In practice, teams tend to lengthen their stints, in hold if the is a yellow flag that allows them to make pitstops in less time.
- Equation 4.42 is to consider the case where there is not a yellow flag during the current lap. Continuing with the previous equation, this is to lengthen the first stint as much as possible.

Since we now have a way to optimize the best running profile for a certain lap of the race, we would use it to evaluate the best strategy for different laps of the race. Thus, we would be updating it every 5 laps.

4.3. Evaluation of the method for different laps of the race

4.3.1. Optimizing race strategies on-line

From the previous section now we have a clear picture of the performance of the different types of tires. Thus, we can start optimizing the race strategy at any lap of the race considering the data until the current lap, and the prior. Let us remember that we are working with the Hungarian Grand Prix from 2022, the objective now is to see how the tire strategy changes as the race progresses. In particular, every 5 laps of the race we evaluate the optimal race strategy for the remaining laps, fixing past decisions from previous laps.

Figure 4.2 shows the evolution of a driver's strategy, as the race progresses. On the x-axis, we have the race lap, and on the y-axis, we have the lap in which the strategy is being evaluated. In each of the rows, we have a marker ($\triangleright$) that allows us to identify the lap in which the driver is evaluating the optimal race strategy. Yellow vertical rectangles in the background represent laps in which there was a yellow flag. As we can see, on the first 3 laps of the race, and the last 4, we have a Yellow Flag.
Let us consider that we are evaluating the strategy on lap 21 of Figure 4.2. This means that we have race times up to lap 20, in addition to the prior information. We can see that at this point, the Soft tire was used for 16 laps, and we reached 4 laps with the Medium tire. From here, the strategy suggests using the Medium tire for an additional 23 laps before making a pit stop and switching to another Medium tire, which will be used until the end of the race, covering 26 laps.

If we look at Figure 4.2, we see the evolution and the importance of the strategy for Max Verstappen, the winner of the 2022 Hungarian Grand Prix. As mentioned in Sportskeeda F1, "Max Verstappen credits Red Bull's 'strategy' after winning 2022 F1 Hungarian GP" [23]. At the start of the first lap, i.e., the bottom row of Figure 4.2, we are assuming the driver is starting with the actual tire compound used at the start of the race (Soft). We observe that the proposed strategy suggests making 2 stints with the Hard tire, and using the Soft for only 1 lap. The reason for suggesting such an early stop is that the optimization model is based on the regression coefficients calibrated with the last Hungarian GP (since by lap 1 there is no information on the 2022 version). Specifically, the Soft compound had a poor performance in the previous race since this was used by most drivers at the beginning of the race, which dried during the course of the race, improving lap time for the other tire compounds.[9]

This effect, the under-rating of the Soft compound, is more pronounced in the first laps of the race, and it vanishes as more lap time data is gathered as more laps are observed. (More discussion of this is provided below after Figure 4.5). Indeed, we observe from Figure 4.2

[9] In the 2020 Hungarian Grand Prix, rain started falling before the race began, but track conditions evolved during the competition. The initial rain led to many drivers starting with wet tires, but as the race progressed, the track began to dry. This resulted in a series of tire changes as some drivers opted for dry (slick) tires to take advantage of the drier track and achieve better performance. As the track was gradually drying out, most drivers initially opted for Soft tires and later switched to Hard tires as the race progressed.

that when the current lap is 6, 11, and 16; the optimization suggests stopping immediately for a Medium compound. Finally, the strategy ended with a first stint of 16 laps with one tire. Soft, and then 2 stints with a Medium tire, the first of 22 laps, and the second of 31 laps.

As said before, we can see the effect of the use of the prior on the regression. In the race from the previous year, the Soft tire showed inferior performance, causing that during the first stages of the race, its use in the strategy was very short. However, as the race being analyzed progresses, additional data becomes available.

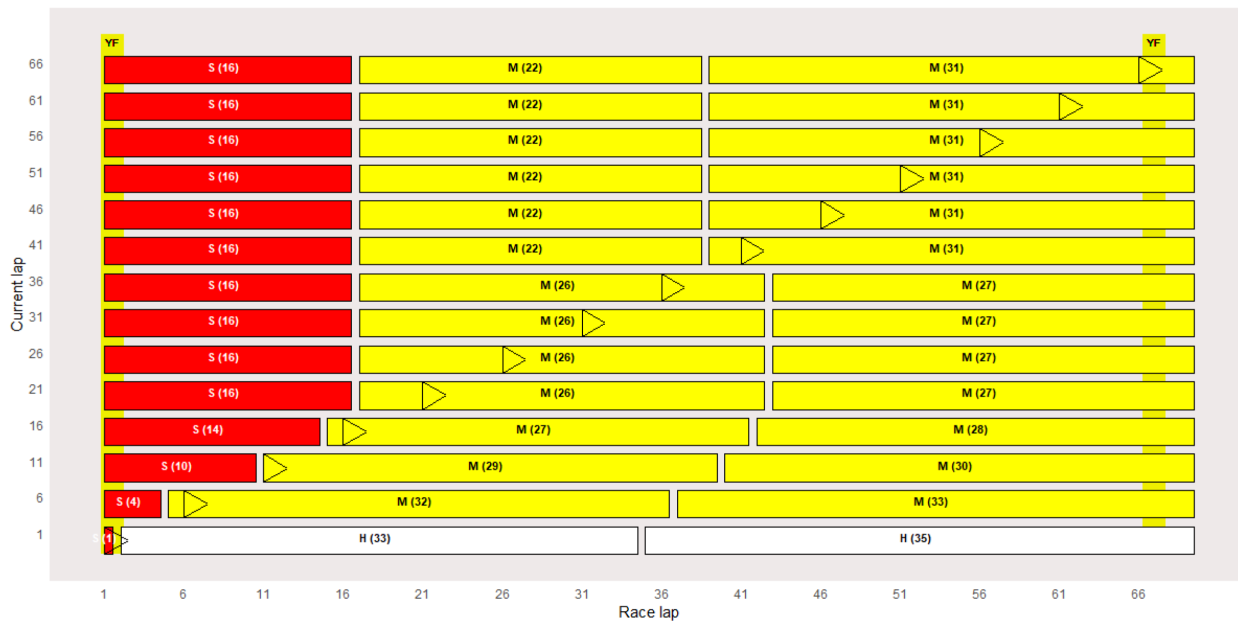

Figure 4.2: Evolution of the race strategy for Max Verstappen for the Hungarian Grand Prix 2022.

In the same way, we now want to analyze another driver's strategy. Let us take a look at Figure 4.3, the strategy for Charles Leclerc. He started the race in third place and finished it in 6th. In Figure 4.3 we can see, for the first 11 laps, we have the same effect as the case before. The strategy suggests using the Hard tire because it performed exceptionally well in the race from the previous year. But as the race progresses, it is no longer considered. During lap 40, Charles Leclerc changed from a Medium to a Hard compound, a controversial action given the information by that lap on the performance of the latter compound. As seen on Sky Sports F1, "Charles Leclerc calls Ferrari strategy 'a disaster' after dropping from first to sixth at Hungarian Grand Prix" (Sky Sports F1, 2023 [22]).
To compensate for the previous decision, the Ferrari team decided to pit him after 15 laps with the Hard tire, in order to run the last 15 laps with a Soft tire. In the end, the strategy considers 4 stints, the first 2 with a Medium tire, then a Hard tire, and finally, a Soft tire.

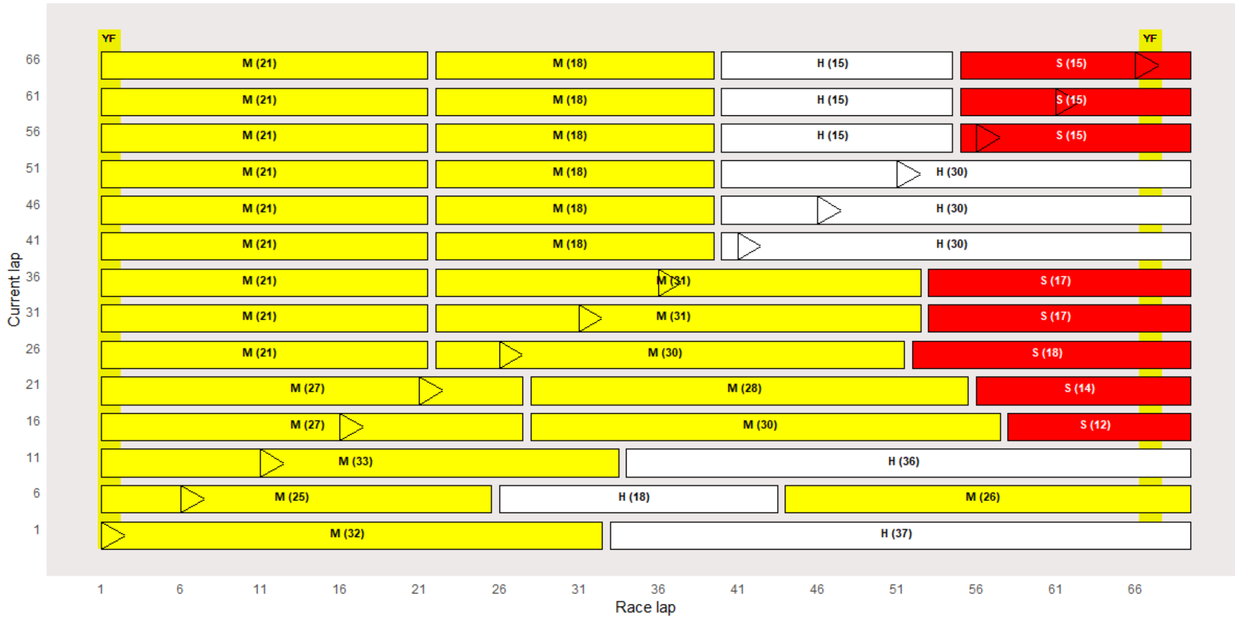

Figure 4.3: Evolution of the race strategy for Charles Leclerc for the Hungarian Grand Prix 2022.

Lastly, let us consider Albon's strategy. He did not have the best tire strategy during the race, because he ended up using a Soft tire compound for more laps than a Medium tire compound. Now, let us take a look at the evolution of his race strategy. As we can see in Figure 4.4, we mostly suggest the use of the Medium tire compound and a short stint with the Soft tire compound at the beginning. On the contrary, Albon ended up making 2 stints of 19 laps with the Medium tire, 1 short with the Soft tire at the beginning, and then another long stint with the Soft tire at the end.

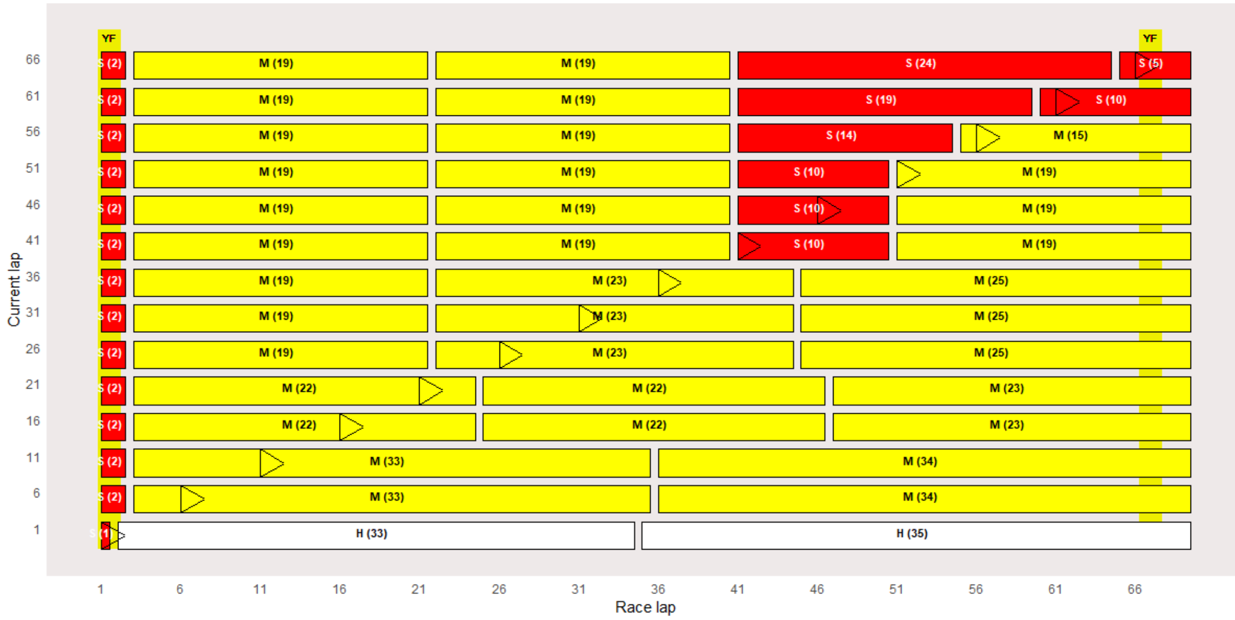

Figure 4.4: Evolution of the race strategy for Alexander Albon for the Hungarian Grand Prix 2022.

4.3.2. Comparison with offline optimal strategy

Let us start by defining online and offline analysis. **Online** refers to the race strategy evaluation with the data available up to the moment we are evaluating it, considering regression calibrated in the same GP race of a previous year. On the other hand, **offline** refers to the evaluation of the strategy at a certain point, but with all the data available. This involves using a regression with all the lap times of the current Grand Prix. The following analyses are performed on the 2022 Hungarian GP.

A visualization of the online and offline cases is depicted in Figure 4.5 for Charles Leclerc. The right panel presents the optimal race strategy for different current laps when (hypothetically) considering all 2022 race data upfront. For example, in lap 5, we are considering that we have all the 70 laps from the race to make a strategy. The objective of doing this is to compare the "ideal" strategy, with the one we are predicting. The left panel illustrates the race strategy considering the previous race as prior and the data until the current lap (as in Figures 4.2, 4.3, and 4.4).

In Figure 4.5 we can see that the main differences, between using and not using the prior, are in the first stages of the race. From lap 15 the strategy is almost the same for both scenarios. The inclusion of the previous year's race data as a prior induces the consideration of the Hard tire at the beginning of the race, due to its impressive performance in that race. In reality, as we can see on the left panel, because that tire has a very poor performance, it is not considered in the strategy. It's worth noting that starting from lap 16, the strategy of the online setting (left panel) aligns very closely with the offline strategy (right panel). The analogous plots of Figure 4.5 for Max Verstappen and Alex Albon are shown in Appendix A

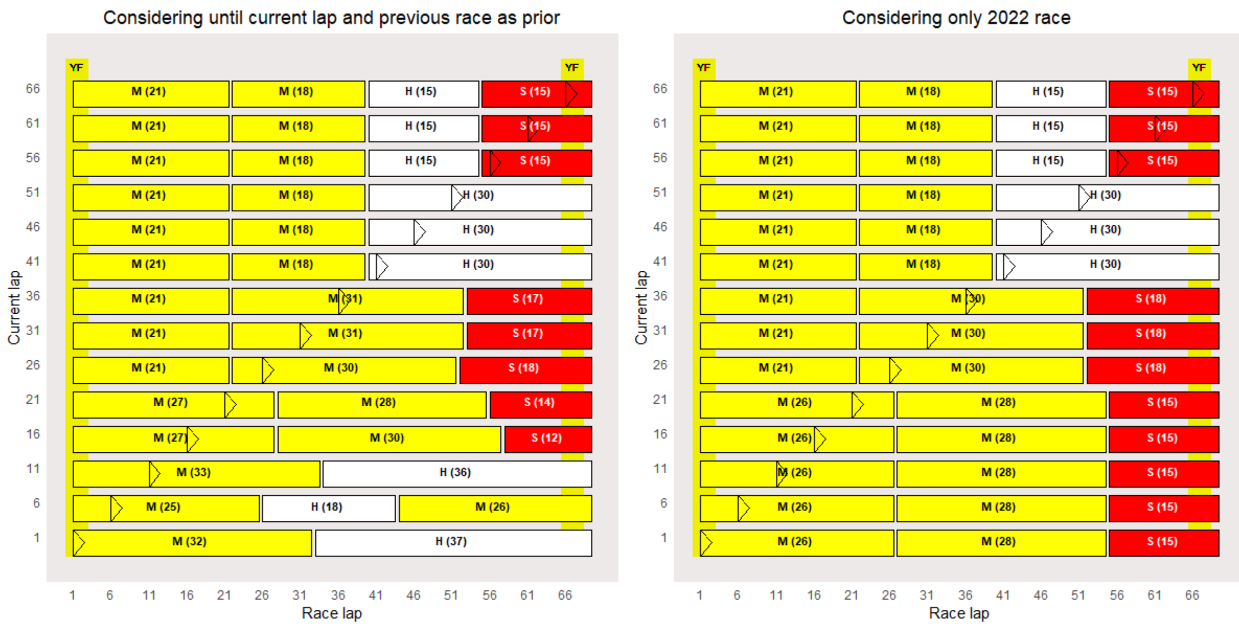

Figure 4.5: Comparison of the evolution of race strategy for Charles Leclerc comparing the scenario where the prior race is taken into account and where it is not considered.

4.3.3. Analysis of all driver strategies

Continuing our analysis, we aim to evaluate how our online strategy compares to the optimal one. In Figure 4.6, we illustrate the time difference between our online proposed strategy and the offline optimal strategy, with lap times calculated using the estimated coefficients from the offline case (i.e., the regression calibrated solely with 2022 lap time data without any prior information). Notably, for Albon and Verstappen, our proposed online strategy closely aligns with the optimal one (offline) from the early stages of the race. However, for Leclerc, the beginning of the race presents more challenges. This discrepancy arises because the online strategy suggests two sets of Hard tires for Leclerc in lap 1, whereas, as mentioned earlier, these tires are outperformed by the other two compounds. Conversely, the Hard compound is not recommended for Verstappen or Albon, possibly due to their initial use of Soft rather than Medium tires.

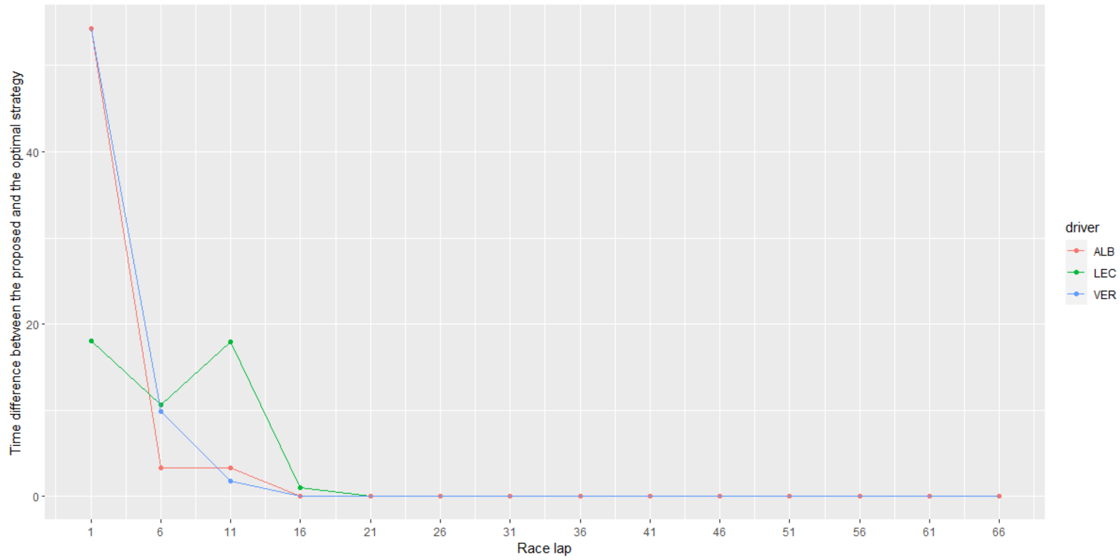

Figure 4.6: Time difference between our proposed strategy and the optimal strategy for Albon, Leclerc, and Verstappen during the Hungarian Grand Prix 2022, evaluating every 5 laps.

In Figure 4.7, we conduct a similar analysis for each driver on the grid. It is noteworthy that our strategy becomes increasingly accurate as the race progresses, with a difference of no more than 2.5 seconds compared to the optimal one for most drivers. It is interesting to note that since lap 20, the strategy of the online setting aligns very closely with the offline strategy. Towards the end of the race, we observe that, for the majority of drivers, the time difference converges to zero.

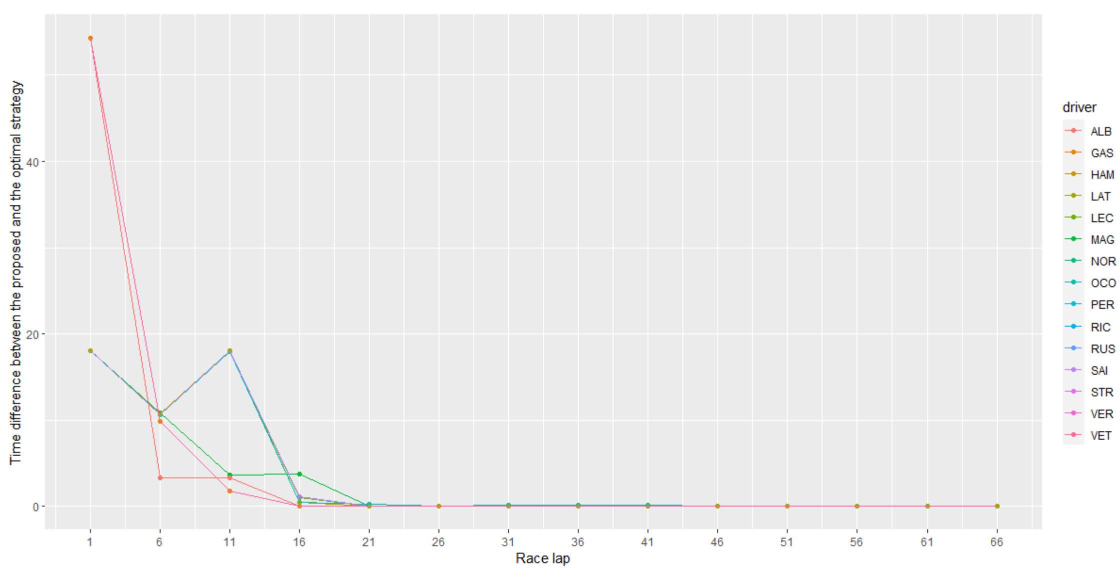

Figure 4.7: Time difference between our proposed strategy and the optimal strategy for every driver during the Hungarian Grand Prix 2022, evaluating every 5 laps.

Let us finish by comparing our strategy with the one the drivers ended up using during the Grand Prix. In Figure 4.8, on the left panel, we show the time difference between the real strategy used by the driver and the optimal strategy we propose when using a regression with all the time laps of the Hungarian GP of 2022, i.e., the offline strategy (x-axis), for every driver (y-axis). On the right panel of Figure 4.8, we show the strategy used by each driver (y-axis) for the race (x-axis). The bottom row of Figure 4.8 corresponds to the optimal strategy. It is worth noticing that other permutations of the stints of this strategy result in the same race time. Let us recall that the time difference is the additional time compared to using a new Soft tire compound in every lap.

As we can see, on the one hand, for drivers like Verstappen, Russell, Hamilton, Vettel, and Perez, our predicted strategy is very similar to the one used in the race. On the other hand, as mentioned before, we have Leclerc's race, a driver who did not choose the best strategy. In fact, our strategy was 18 seconds faster than the strategy used by Leclerc. Finally, let us consider Magnussen's race, the driver that had the greatest difference in time between the real strategy and the one we propose. This is, if he had used our strategy, he might have ended the race almost 25 seconds before. We can see that the strategy he used, is very different from the optimal one, considering a long stint with a Hard tire, and starting the race with a Medium tire compound.

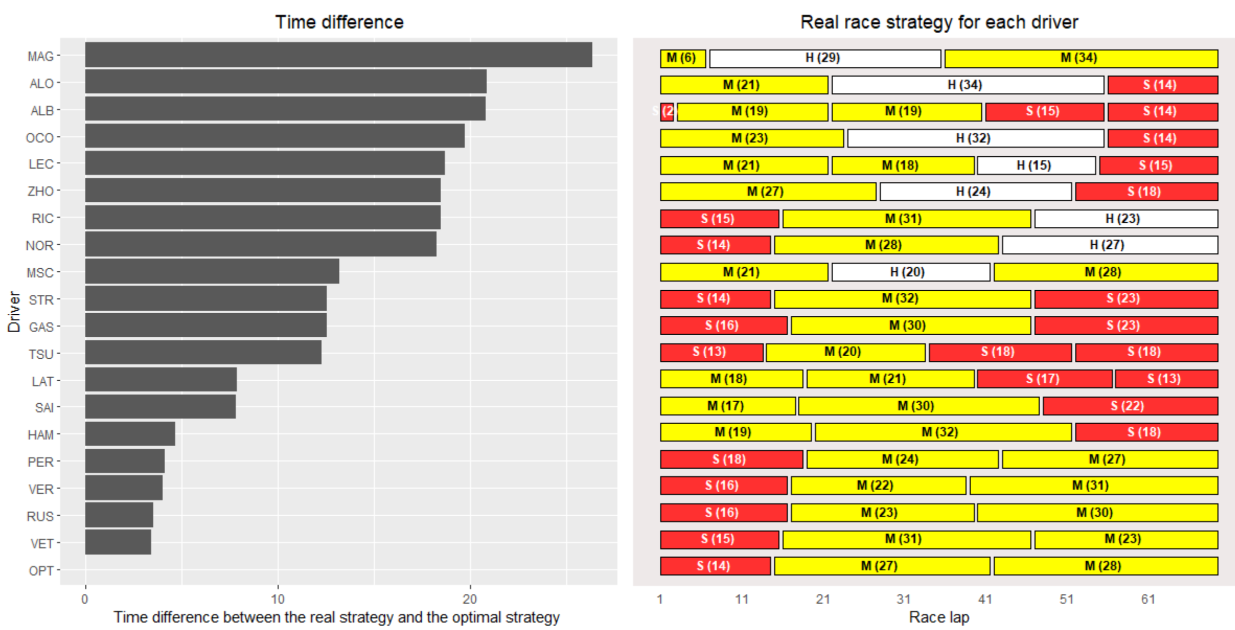

Figure 4.8: Time difference between the real strategy used by the driver and the optimal strategy we propose for every driver during the Hungarian Grand Prix 2022.

From the perspective of tire strategy, we can assert that starting a race with Medium compound tires is recommended. This choice strikes the best balance between tire wear and performance, allowing for a reasonable time window to gather more race data in order to have a more reliable optimized race strategy. Conversely, the Soft tire, due to its rapid wear, affords very few laps to formulate a strategy, while the Hard tire carries a higher risk of poor performance.

In Figure 4.9, we show the time difference between the real strategy used by the driver and the optimal strategy we propose when using a regression with all the time laps of the Hungarian GP of 2022 (x-axis), for every driver (y-axis). We are also showing on the x-axis the effect of the team obtained from the regression. It is also important to notice that the drivers are ordered according to their finishing position during the Grand Prix. As we can see in Figure 4.9, there are some drivers, like Tsunoda, that even though they had a good strategy, the team effect made their time increase. This is because even though a driver has a good strategy, their skills and the car can make a difference.

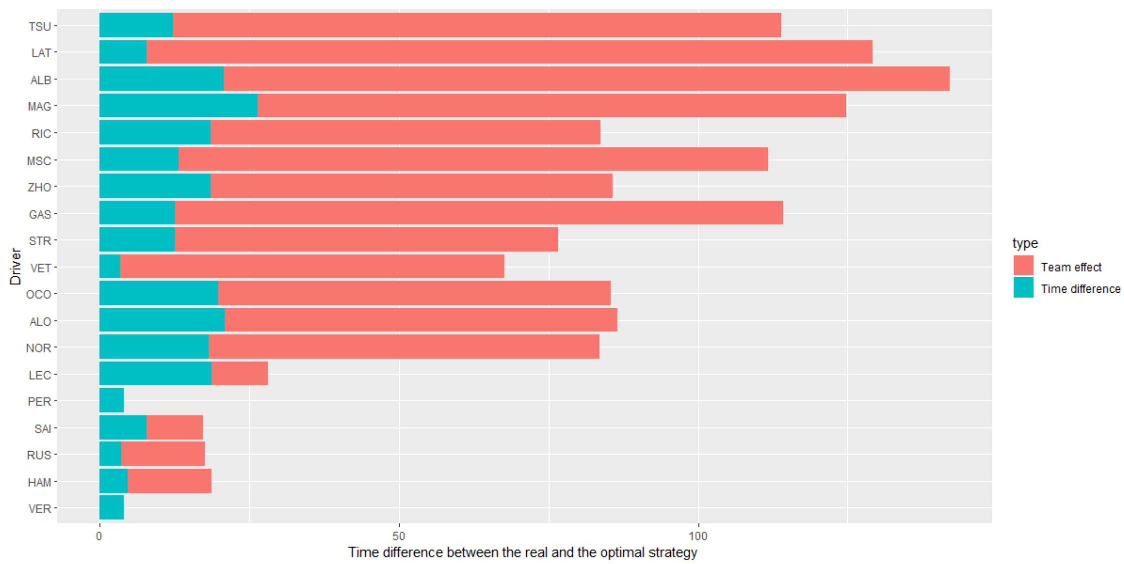

Figure 4.9: Time difference between the real strategy used by the driver and the optimal strategy we propose for every driver during the Hungarian Grand Prix 2022, considering the team effect.

Chapter 5

Conclusions

In this work, we utilized the dataset “Formula 1 World Championship (1950-2020)” by Rao, Rohan, obtained from Kaggle [21]. This dataset encompasses information on Formula 1 races, drivers, constructors, qualifying sessions, circuits, lap times, pit stops, and championships from 1950 through the latest 2023 season. Additionally, to gather information for constructing the dataset concerning stints, we relied on the Racefans[19] website, which provided comprehensive data on pit stops for every race.

From the data obtained from this dataset, we employed regression analysis to investigate the impact of tire wear on lap times for different tire compounds. Since there are non-observable factors in the data that lead to predictions with high errors, we introduced two methodologies for removing outliers. These methodologies involve iterations by calibrating a regression and eliminating data points whose residuals are flagged as outliers. On one hand, we implemented the *Interquartile* method, which is based on percentiles and data distribution. On the other hand, we introduced the *Confidence Region* method, which imposes a threshold on the probability of obtaining a prediction as unusual as the one obtained. We opted to work with the *Interquartile* method since it retains more data and still allows us to make reasonable predictions.

Through the management of data outliers using the *Interquartile* method, we derived coefficients associated with various factors influencing a driver’s lap times performance during the race. Subsequently, by utilizing these coefficients related to different tire types, we estimated their performance throughout the race. Given that our objective was to provide the optimal strategy as the race progresses, a challenge arose in the early stages of the race, where there was insufficient data to make reliable predictions. To address this, we employed data from the previous year’s race as prior information. These prior beliefs were then combined with observed data to update our knowledge about the parameters, resulting in a posterior distribution. Using the estimated coefficients of (i)the tire compounds, (ii) lap wear across tire compounds, and (iii)the time loss incurred during pit stops, we optimized the race strategy for the remaining portion of the race. This optimization was achieved by utilizing an MIQP (Mixed-Integer Quadratic Programming) formulation, which determined the length and tire compound for all future stints. Notably, the MIQP was solved within a fraction of a second. Furthermore, we provided a non-trivial upper bound for the maximum number of remaining pit stops for the remainder of the race. This upper bound was employed by the MIQP as the number of decision variables is linearly related to this number.

The methods elucidated were applied to several races of the 2022 season, encompassing both drivers whose pit stop strategies raised questions and those whose strategies were successful. It is noteworthy that the results obtained, in particular, race strategies, differ in some cases from the actual race strategies used by some drivers. In cases where, for instance, a Soft tire was used for more laps than a Medium tire, the solution we proposed varied. It is also imperative to mention that the method's suggestions in the first 15 laps may not be as reliable. Due to the heavier influence of prior weights at the beginning of the race, when data availability is limited, results obtained during these early stages are less accurate. On the contrary, it is interesting to observe that as the race progresses, the suggested online optimal race strategies resemble the offline optimal strategy. Another important conclusion is that, although strategy plays a fundamental role in determining race outcomes, the driver, their experience, and the car are also of paramount importance. Even with an impeccable strategy, a driver's style, or the car, can negatively impact performance, resulting in longer race times than initially estimated for the strategy.

In conclusion, we successfully achieved our primary goal: optimizing tire strategy during an F1 race and adapting it as the race unfolds. It is worth emphasizing that while our focus centered on Formula 1, this work can be replicated in any motorsport category.

In future works, this study could be extended. On one hand, the climate could be taken into consideration, which would involve considering rain tires. On the other hand, yellow flags could be taken into account. Unlike this study, the probability of a yellow flag occurring could be added. This could be done with the aim of making the work more realistic.

Bibliography

[1] Agathangelou, B., & Gascoyne, M. (1998). Aerodynamic design considerations of a Formula 1 racing car. SAE International, 724, 1–4.

[2] Bell, A., Smith, J., Sabel, C., & Jones, K. (2016). Formula for success: Multilevel modeling of Formula One Driver and Constructor performance, 1950–2014. Journal of Quantitative Analysis in Sports, 12(2), 99-112. doi:10.1515/jqas-2015-0050

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

[4] Bloem, B. R., Valkenburg, V. V., Slabbekoorn, M., & Willemsen, M. D. (2001). The Multiple Tasks Test: Development and normal strategies. Gait & Posture, 14(3), 191-202. doi:10.1016/S0966-6362(01)00141-2

[5] Borsboom, O., Fahdzyana, C. A., Hofman, T., & Salazar, M. (2021). A convex optimization framework for minimum lap time design and control of electric race cars. IEEE Transactions on Vehicular Technology, 70(9), 8478–8489. IEEE.

[6] Choo, C. L. W. (2015). Real-time decision making in motorsports: analytics for improving professional car race strategy. Massachusetts Institute of Technology.

[7] Collantine, K. (n.d.). F1 tyres: Changes, compounds, rules, colours, blankets. GPFans. Retrieved from https://www.gpfans.com/en/f1-news/100347/f1-tyres-changes-compounds-rules-colours-blankets/

[8] Duhr, P., Buccheri, D., Balerna, C., Cerofolini, A., & Onder, C. H. (2023). Minimum-Race-Time Energy Allocation Strategies for the Hybrid-Electric Formula 1 Power Unit. IEEE Transactions on Vehicular Technology. doi:10.1109/TVT.2023.3237388

[9] Elizalde, P. (2022, May 28). The many factors that determine the fate of an F1 title. Motorsport Week. Retrieved from https://www.motorsportweek.com/2022/05/28/the-many-factors-that-determine-the-fate-of-an-f1-title/

[10] Formula 1. (Year of Access). Formula 1 Official Website. Retrieved from https://www.formula1.com/

[11] FormulaPedia. (n.d.). Formula One Tires. Retrieved from https://formulapedia.com/formula-one-tires/

[12] Greasley, A., Panchal, G., & Samvedi, A. (2022). The Use of Simulation with Machine Learning and Optimization for a Digital Twin-A Case on Formula 1 DSS. In 2022 Winter Simulation Conference (WSC) (pp. 2198–2209). IEEE.

[13] Heilmeier, A., Graf, M., & Lienkamp, M. (2018). A Race Simulation for Strategy Decisions in Circuit Motorsports. In 2018 21st International Conference on Intelligent Transportation Systems (ITSC) (pp. 2986-2993). doi:10.1109/ITSC.2018.8570012

[14] Heilmeier, A., Thomaser, A., Graf, M., & Betz, J. (2020). Virtual strategy engineer: Using artificial neural networks for making race strategy decisions in circuit motorsport. Applied Sciences, 10(21), 7805. MDPI.

[15] Heine, O. F. C., & Thraves, C. (2023). On the optimization of pit stop strategies via dynamic programming. Central European Journal of Operations Research, 31(1), 239–268. Springer.

[16] Hu, T. (2023). Analysis of The Venturi Tunnel and Ground Effect. Highlights in Science, Engineering and Technology, 38, 695-698. doi:10.54097/hset.v38i.5933

[17] Jin, M. H., & Yoo, T. S. (2017). A study on the influence of the safety car on Formula 1 racing. CORE. Retrieved from https://core.ac.uk/reader/83928830

[18] Limebeer, D. J. N., Perantoni, G., & Rao, A. V. (2014). Optimal control of formula one car energy recovery systems. International Journal of Control, 87(10), 2065–2080. Taylor & Francis.

[19] RaceFans. Retrieved from https://www.racefans.net/

[20] Pirelli. Formula 1® Tires. Retrieved from https://www.pirelli.com/tires/en-us/motorsport/f1/tires

[21] Rao, R. Formula 1 World Championship (1950-2020). Kaggle. Retrieved from https://www.kaggle.com/datasets/rohanrao/formula-1-world-championship-1950-2020

[22] Sky Sports F1. (2022). Charles Leclerc calls Ferrari strategy a disaster after dropping from first to sixth at Hungarian Grand Prix. Retrieved from https://www.skysports.com/f1/news/12433/12662820/charles-leclerc-calls-ferrari-strategy-a-disaster-after-dropping-from-first-to-sixth-at-hungarian-grand-prix

[23] Sportskeeda. (2022). Max Verstappen credits Red Bull's strategy for winning the 2022 F1 Hungarian GP. Retrieved from https://www.sportskeeda.com/f1/news-max-verstappen-credits-red-bull-s-strategy-winning-2022-f1-hungarian-gp

[24] Talbi, N. (2015). Optimization of Formula 1 Racing.

[25] Tremlett, A. J., & Limebeer, D. J. N. (2016). Optimal tyre usage for a Formula One car. Vehicle System Dynamics, 54(10), 1448-1473. doi:10.1080/00423114.2016.1213861

[26] Tulabandhula, T., & Rudin, C. (2014). Tire changes, fresh air, and yellow flags: challenges in predictive analytics for professional racing. Big data, 2(2), 97-112. Mary Ann Liebert, Inc. 140 Huguenot Street, 3rd Floor New Rochelle, NY 10801 USA.

[27] West, W. J., & Limebeer, D. J. N. (2020). Optimal tyre management of a formula one car. IFAC-PapersOnLine, 53(2), 14456–14461. Elsevier.

Annex A

Hungarian Grand Prix 2022

Continuing with Max Verstappen's strategy, we will proceed to contrast the evolution of the race strategy, taking into account whether the preceding race is considered or not. The right panel presents the race strategy only based on the 2022 race, while the left panel illustrates the race strategy considering the previous race as prior. It's important to note that in the case where we do not have a prior, we are considering that we have available all the laps from the actual race. For example, in lap 5, we are considering that we have all the 70 laps from the race to make a strategy. The objective of doing this is to be able to compare the "ideal" strategy, with the one we are predicting. In Figure A.1 we can see that the main differences, between using and not using the prior, are in the first stages of the race. It's worth noting that starting from lap 16, the strategy that considers the prior aligns closely with the strategy that focuses solely on the 2022 race.

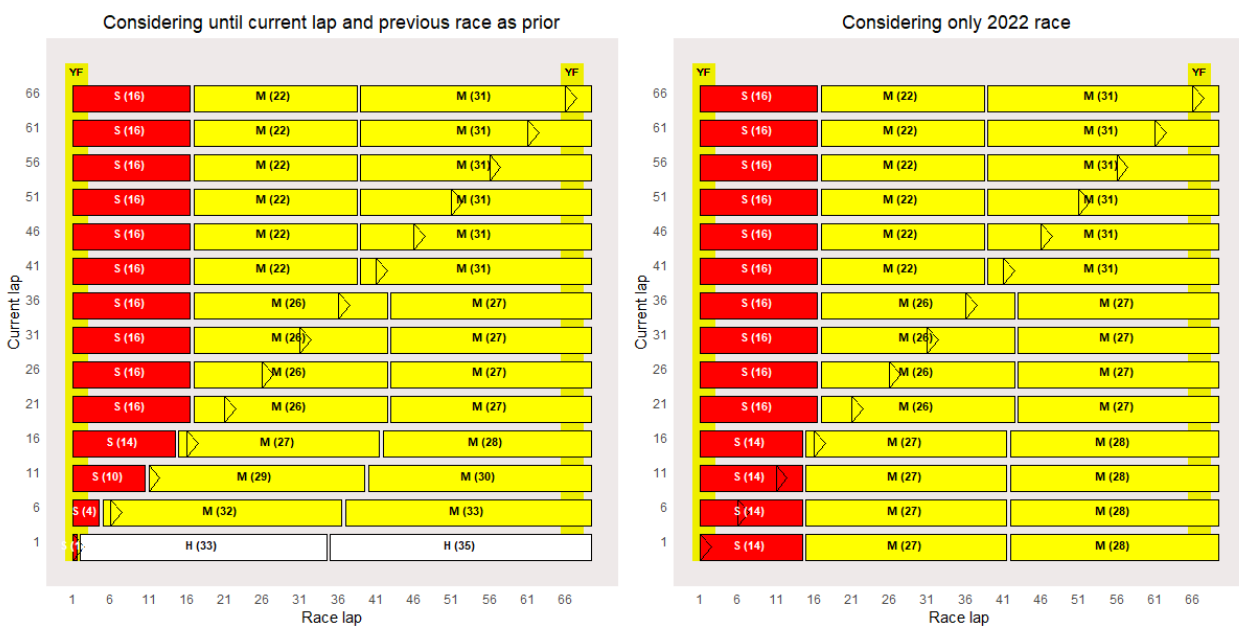

Figure A.1: Comparison of the evolution of race strategy for Max Verstappen comparing the scenario where the prior race is taken into account and where it is not considered.

In Figure A.2 we show the same analysis for Alexander Albon. The inclusion of the previous year's race data as a prior induces that we do not consider the Soft tire, due to its poor performance in that race. In reality, as we can see on the left panel, because that tire has a good performance, it is considered in the strategy. It's worth noting that starting from lap 10, the strategy that considers the prior aligns closely with the strategy that focuses solely on the 2022 race.

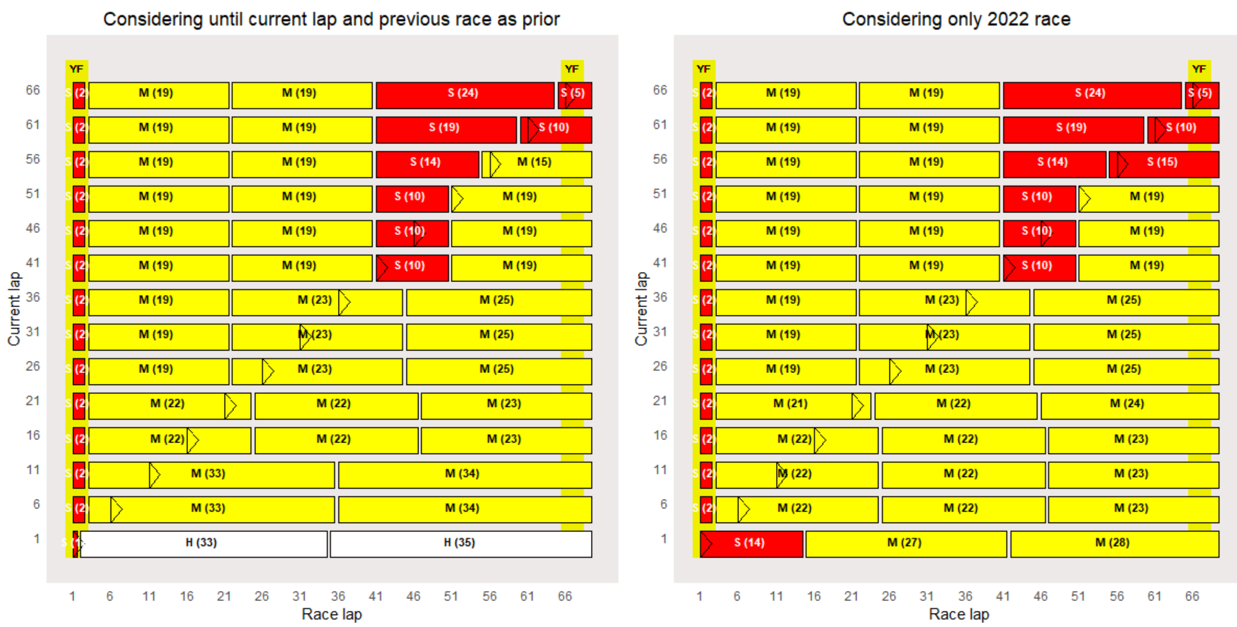

Figure A.2: Comparison of the evolution of race strategy for Alexander Albon comparing the scenario where the prior race is taken into account and where it is not considered.

In Figure a.3, we show the time difference between the real strategy used by the driver and the optimal strategy we propose when using a regression with all the time laps of the Hungarian GP of 2022(x-axis), for every driver (y-axis). We are also showing on the x-axis the effect of the team obtained from the regression. It is also important to notice that the drivers are ordered according to their finishing position during the Grand Prix. As we can see in Figure 4.9, there are some drivers, like Latifi and Albon, that even though they had a good strategy, the team effect made their time increase. This is because even though a team has a good strategy, their car can make a difference.

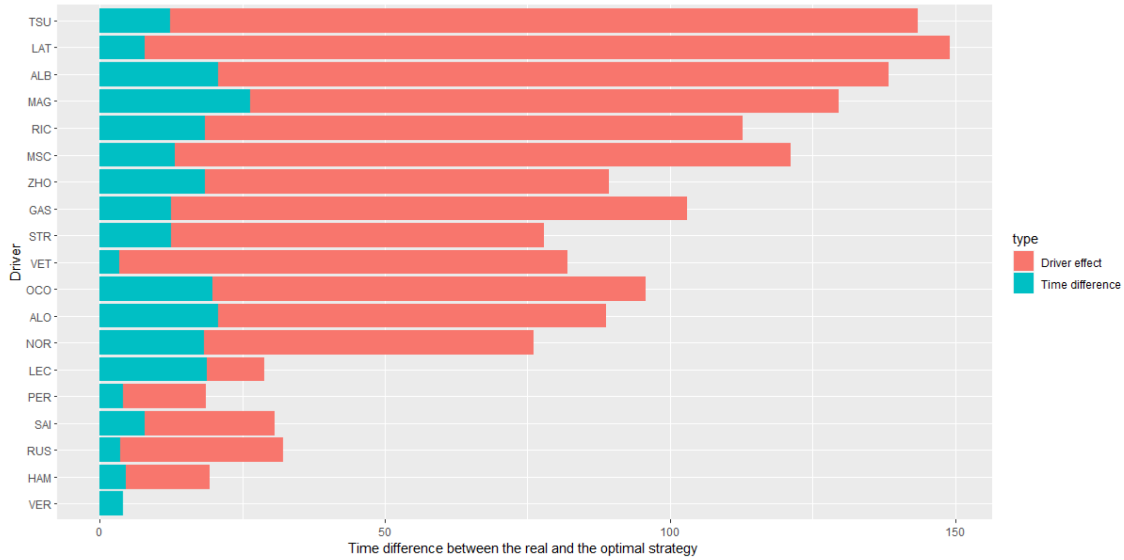

Figure A.3: Time difference between the real strategy used by the driver and the optimal strategy we propose for every driver during the Hungarian Grand Prix 2022.

Annex B

Abu Dhabi Grand Prix 2022

We would make the same analysis with the Abu Dhabi Grand Prix 2022, by applying the Interquartile method to handle the outliers. In Figure A.1 we can see the data that this method eliminates.

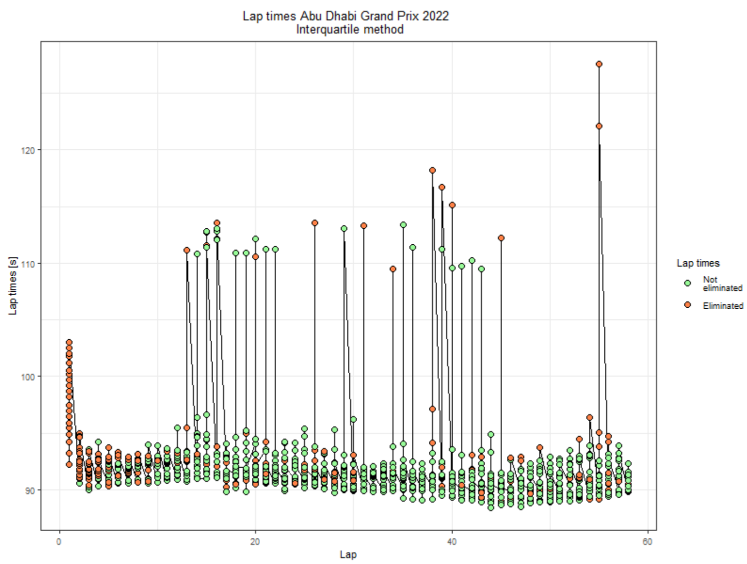

Figure B.1: Lap times Abu Dhabi Grand Prix 2022

In Figure A.2 we can see the difference between considering the data from the previous year in the coefficient prior and not. On the left, we have the tire performance for each compound by considering only data from the first 15 laps of the race. In the image in the center, we can see the tires' performance calculated from the previous year's race data. And, in the image to the right we see again the analysis of the performance of the tires on lap 15 of the race, but this time, considering the previous year's race as prior. When we make predictions about tire performance by only considering data from the first 15 laps of the race, the results, for both methods, are not satisfactory, we have a high standard deviation for all the tires, showing a large amount of variation. As we have more data available, the results improve.

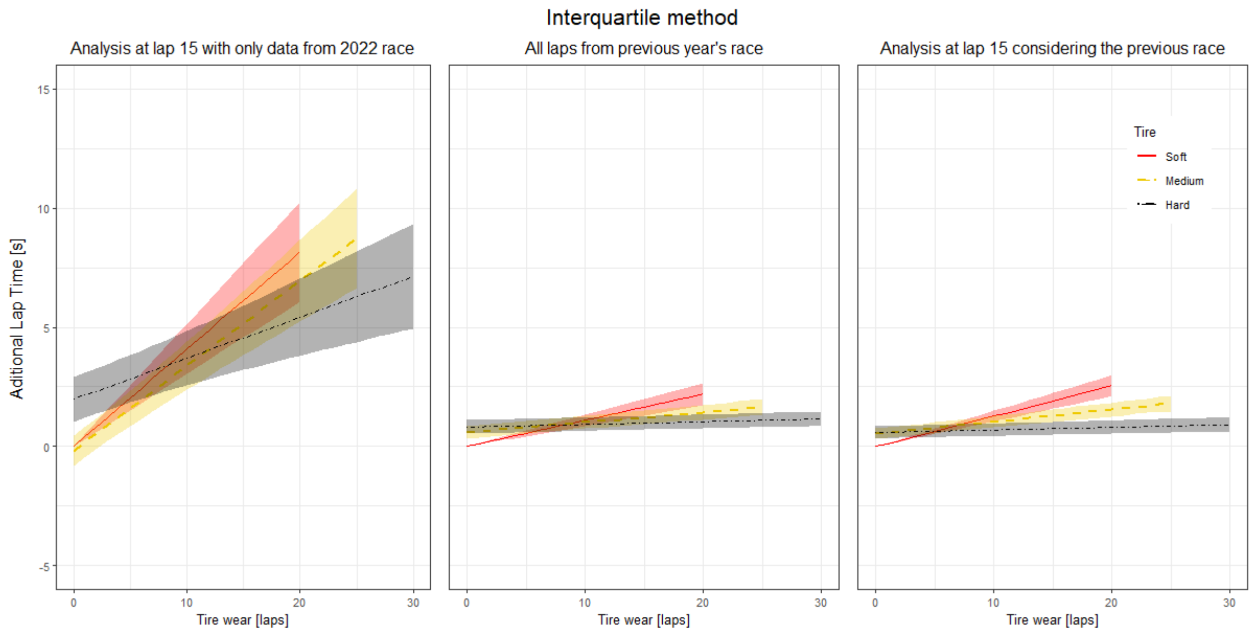

Figure B.2: Comparison of the tire performance at lap 15 Abu Dhabi Grand Prix 2022.

Continuing with the analysis, we can see the tire performance for all the tire types, at the end of the race, and considering data from the previous year's race.

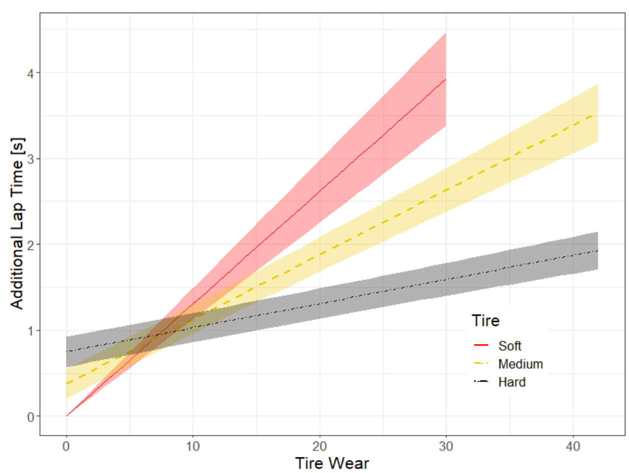

Figure B.3: Tire performance at the end of the Abu Dhabi Grand Prix 2022, considering data from the previous year's race

In the next Figure, we have the evolution of the race strategy for Max Verstappen for the Abu Dhabi Grand Prix 2022. Race that he ended up winning. As we can see, the strategy we suggest is very similar to the one the driver used. In the beginning, we suggested using a Medium tire for 16 laps, and then a Hard tire for 41 laps, and Verstappen ended up using the Medium tire for 20 laps, and the Hard tire for 37 laps.

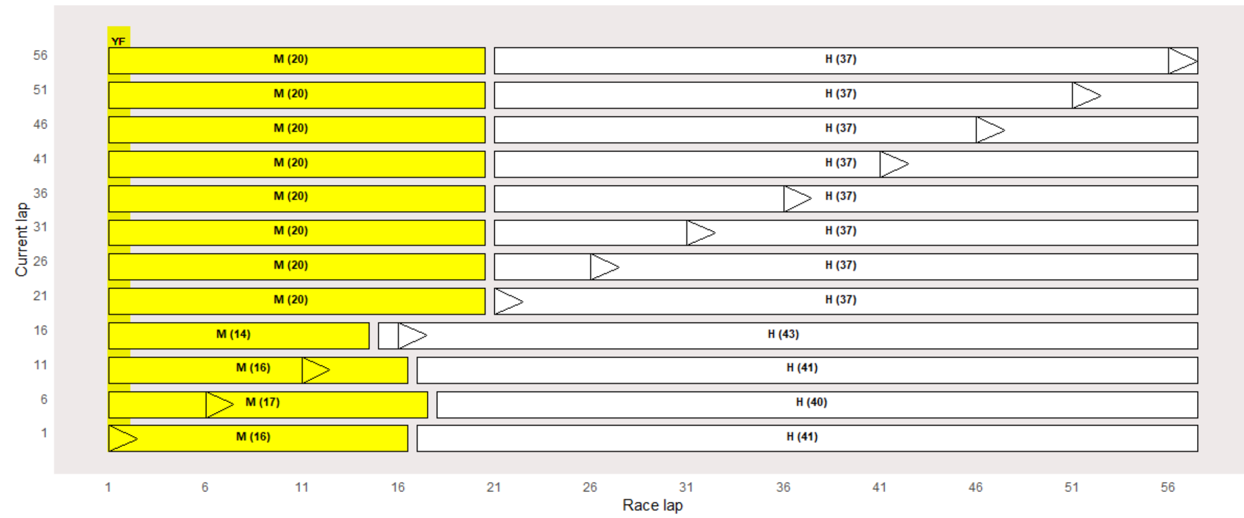

Figure B.4: Evolution of the race strategy for Max Verstappen for the Abu Dhabi Grand Prix 2022.

In Figure B.5 we have the comparison of the evolution of race strategy for Max Verstappen comparing the scenario where the prior race is taken into account and where it is not considered. It is interesting to note that at the first 20 laps, the on-line strategy differs from the off-line. But then, the are very similar.

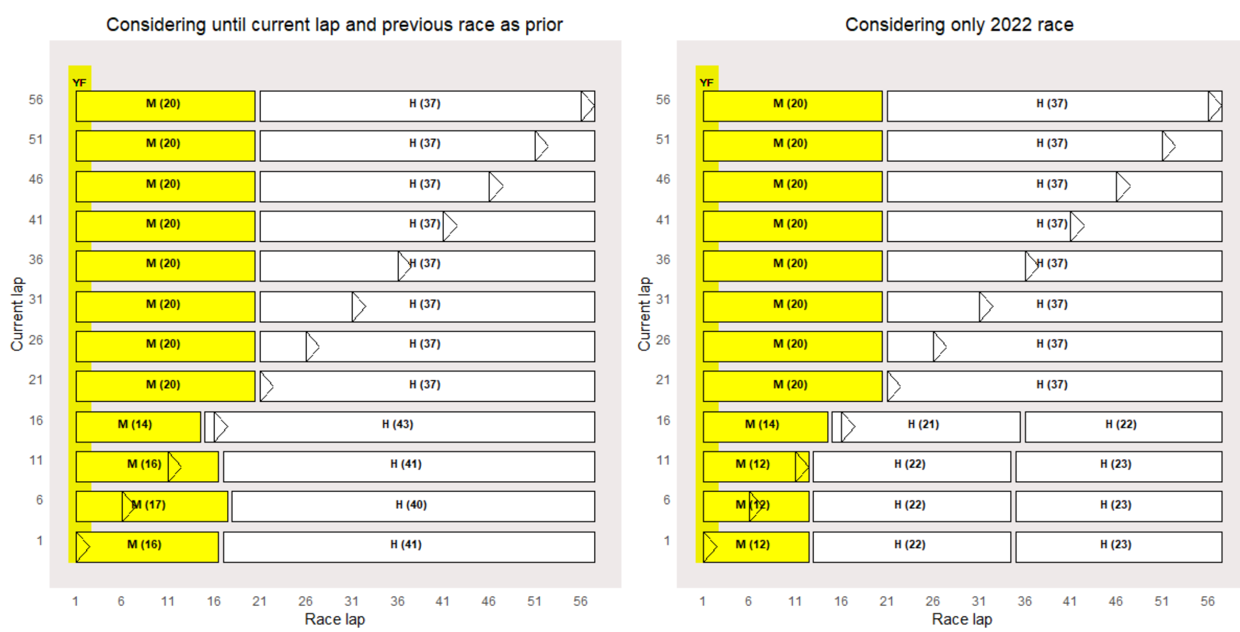

Figure B.5: Comparison of the evolution of race strategy for Max Verstappen comparing the scenario where the prior race is taken into account and where it is not considered for the Abu Dhabi Grand Prix 2022.

For the next part, we wanted to evaluate cases in which drivers employed unconventional strategies. For instance, let us consider the strategy Yuki Tsunoda used at the Abu Dhabi

Grand Prix 2022. He used a Soft tire for 19 laps and a Medium one for 14. What stands out here is that the Medium tire was used for more laps than the Soft tire, even though it wears out faster. In Figure B.6, we suggested doing one pitstop, and only using the Medium and the Hard tire. Instead, on lap 39, the team decided to put on a Soft tire compound until the end of the race. It is also important to notice the differences between using, and not using the race from the previous year as a prior. When we are only considering the 2022 race, we suggest 3 stints, meanwhile, when we consider the prior, we suggest only 2 stints.

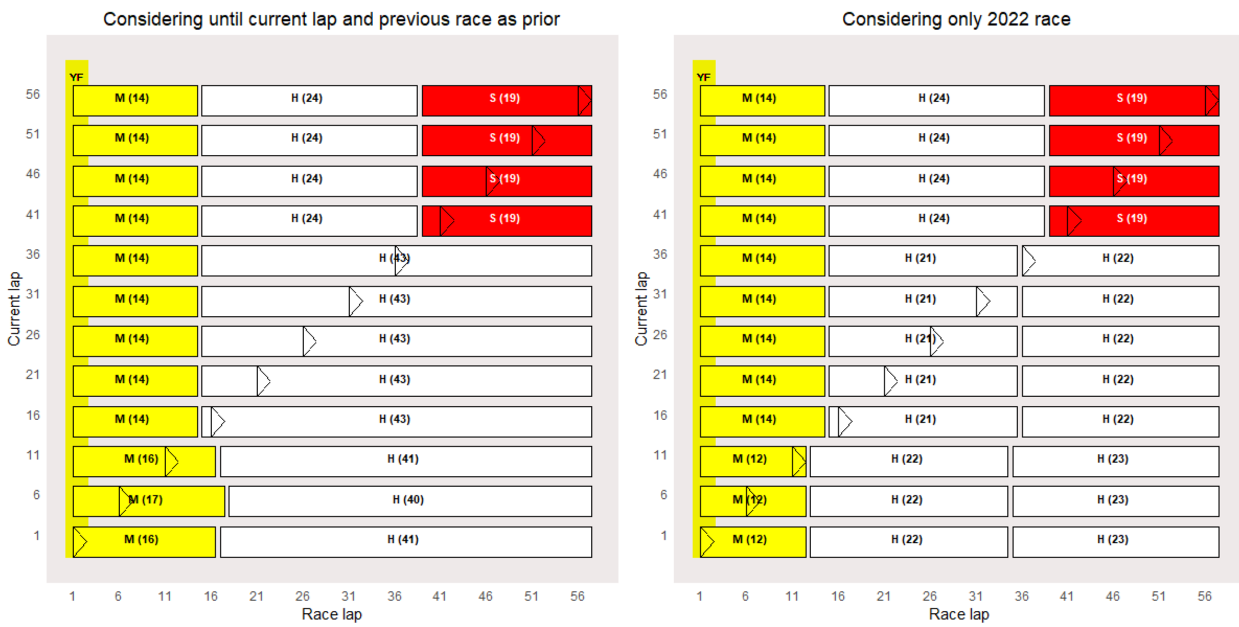

Figure B.6: Comparison of the evolution of race strategy for Yuki Tsunoda comparing the scenario where the prior race is taken into account and where it is not considered for the Abu Dhabi Grand Prix 2022.

Finally, we have the strategy of Mick Schumacher. At the beginning, we suggested using a Medium tire for 16 laps, and then a Hard one for 41 laps. Instead, Schumacher's team decided on lap 35 to put on a Medium tire, which ended up used for more laps than the Hard one. Similarly to the case of Yuki Tsunoda, Mick Schumacher used a tire that wears faster for more laps than the one that lasts more.

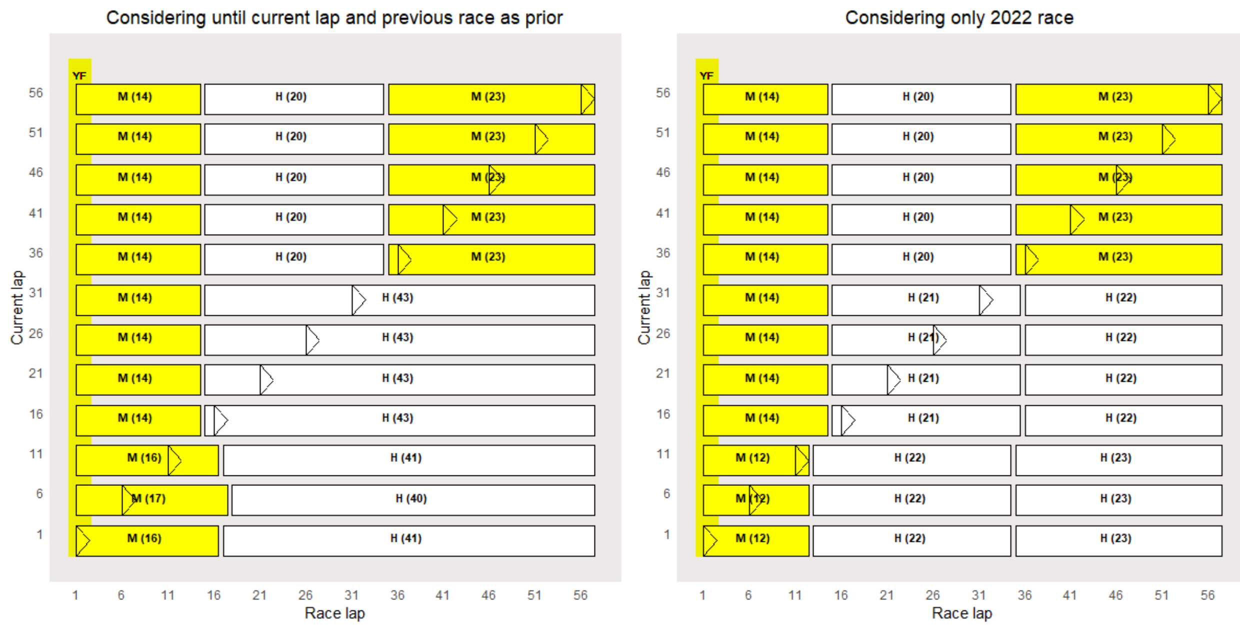

Figure B.7: Comparison of the evolution of race strategy for Mick Schumacher comparing the scenario where the prior race is taken into account and where it is not considered for the Abu Dhabi Grand Prix 2022.

Continuing our analysis, we aim to evaluate how our online strategy compares to the optimal one for each driver of the grid. In Figure B.8, we illustrate the time difference between our online proposed strategy and the offline optimal strategy, with lap times calculated using the estimated coefficients from the offline case (i.e., the regression calibrated solely with 2022 lap time data without any prior information). It is noteworthy that our strategy becomes increasingly accurate as the race progresses, with a difference of no more than 5 seconds compared to the optimal one for most drivers. It is interesting to note that since lap 30, the strategy of the online setting aligns very closely with the offline strategy. Towards the end of the race, we observe that, for the majority of drivers, the time difference converges to zero.

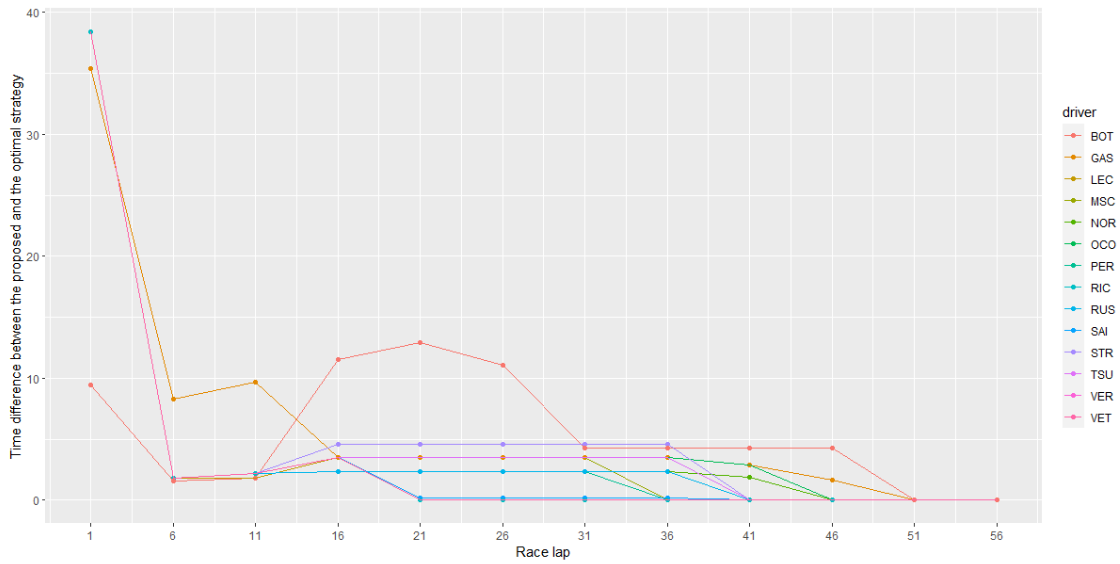

Figure B.8: Time difference between our proposed strategy and the optimal strategy for every driver during the Abu Dhabi Grand Prix 2022, evaluating every 5 laps.

In the figure below, on the left panel, we have the time difference between the real strategy used by the driver and the optimal strategy we propose (x-axis), for every driver (y-axis). On the right panel, we have the strategy used by each driver (y-axis) for the race (x-axis). Let us remember that the time difference is the additional time compared to using a new Soft tire compound in every lap. As we can see, on one hand, for drivers like Perez, Ocon, Stroll, and Norris, our predicted strategy is very similar to the one used in the race. On the other hand, we have Magnussen's race, a driver who didn't choose the best strategy. In fact, our strategy was more than 10 seconds faster than the real strategy. It is also interesting to note that the strategies that considered the Soft tire were among the slower ones.

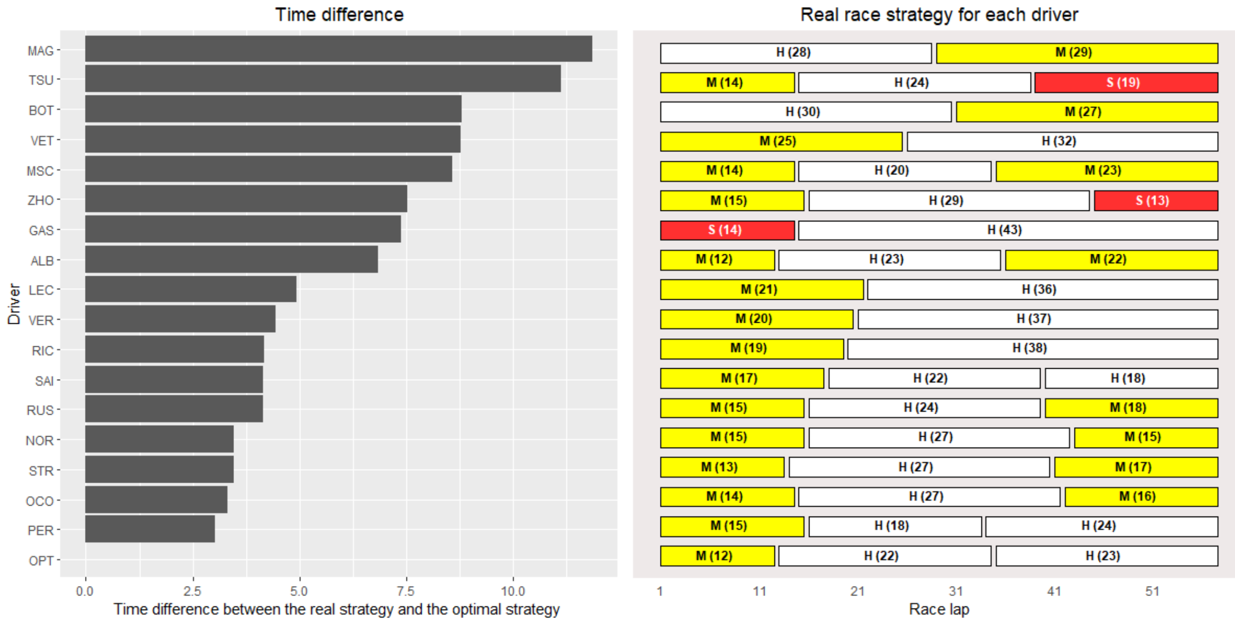

Figure B.9: Time difference between the real strategy used by the driver and the optimal strategy we propose for every driver during the Abu Dhabi Grand Prix 2022.

In Figure B.10, we show the time difference between the real strategy used by the driver and the optimal strategy we propose when using a regression with all the time laps of the Abu Dhabi GP of 2022 (x-axis), for every driver (y-axis). We are also showing on the x-axis the effect of the driver obtained from the regression. It is also important to notice that the drivers are ordered according to their finishing position during the Grand Prix. As we can see in Figure 4.9, there are some drivers, like Norris, that even though they had a good strategy, the driver effect made their time increase. This is because even though a driver has a good strategy, their skills, and the car can make a difference. In Figure B.11 we are making the same analysis, but we are also showing on the x-axis the effect of the team obtained from the regression.

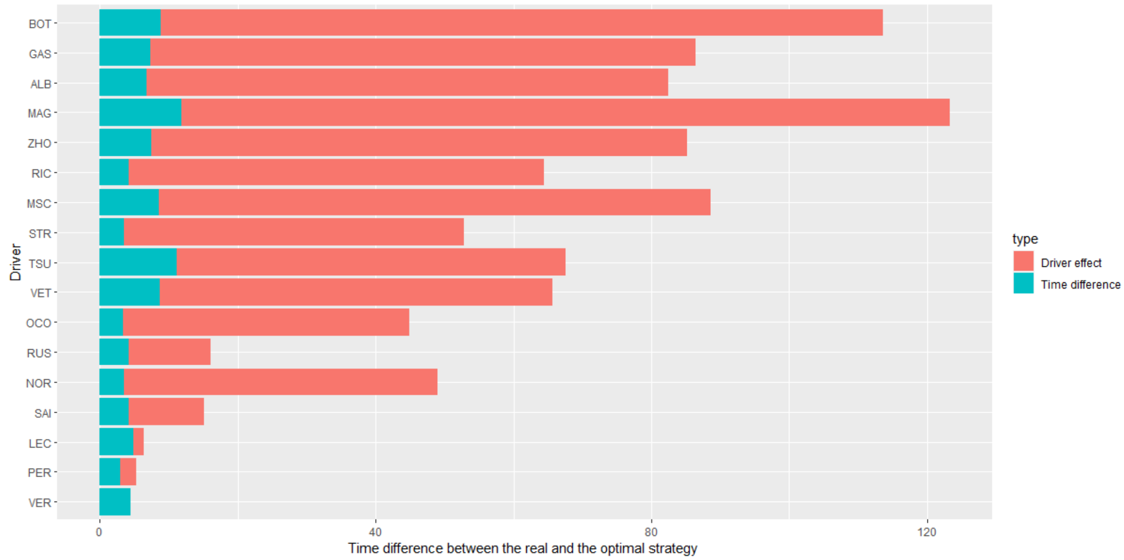

Figure B.10: Time difference between the real strategy used by the driver and the optimal strategy we propose for every driver during the Abu Dhabi Grand Prix 2022.

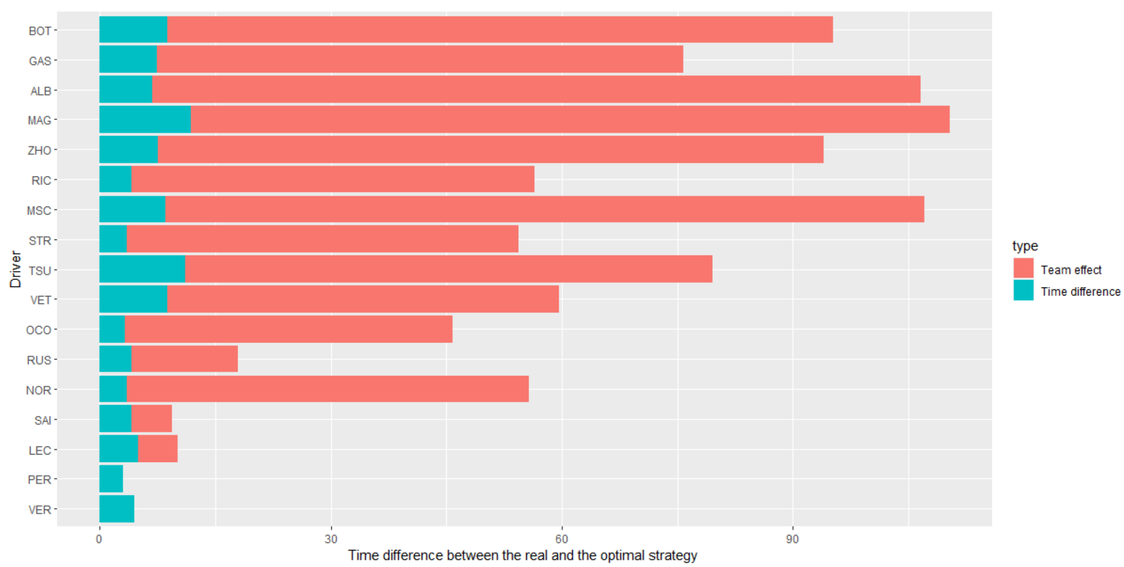

Figure B.11: Time difference between the real strategy used by the driver and the optimal strategy we propose for every driver during the Abu Dhabi Grand Prix 2022.

Annex C

Dutch Grand Prix 2022

We would make the same analysis with the Dutch Grand Prix 2022, by applying the Interquartile method to handle the outliers. In Figure C.1 we can see the data that this method eliminates.

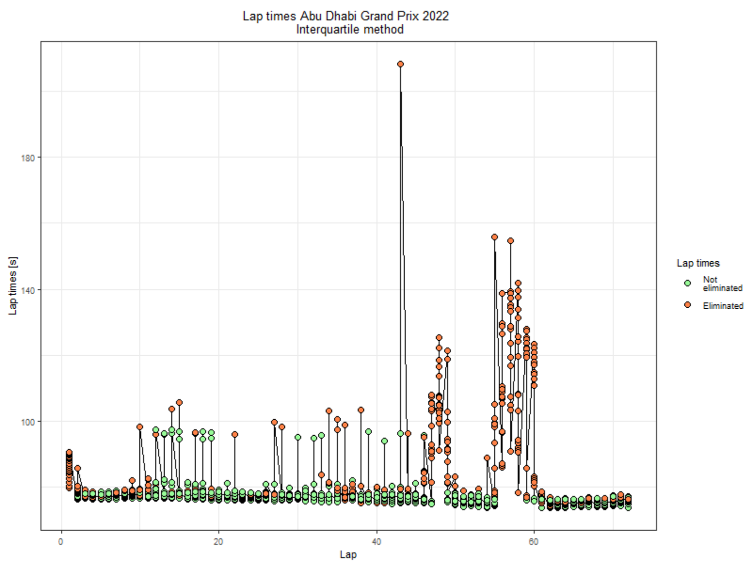

Figure C.1: Lap times Dutch Grand Prix 2022

In Figure C.2 we can see the difference between considering the data from the previous year in the coefficient prior and not. On the left, we have the tire performance for each compound by considering only data from the first 15 laps of the race. In the image in the center, we can see the tires' performance calculated from the previous year's race data. And, in the image to the right we see again the analysis of the performance of the tires on lap 15 of the race, but this time, considering the previous year's race as prior. When we make predictions about tire performance by only considering data from the first 15 laps of the race, the results, for both methods, are not satisfactory, we have a high standard deviation for all the tires, showing a large amount of variation. As we have more data available, the results improve.

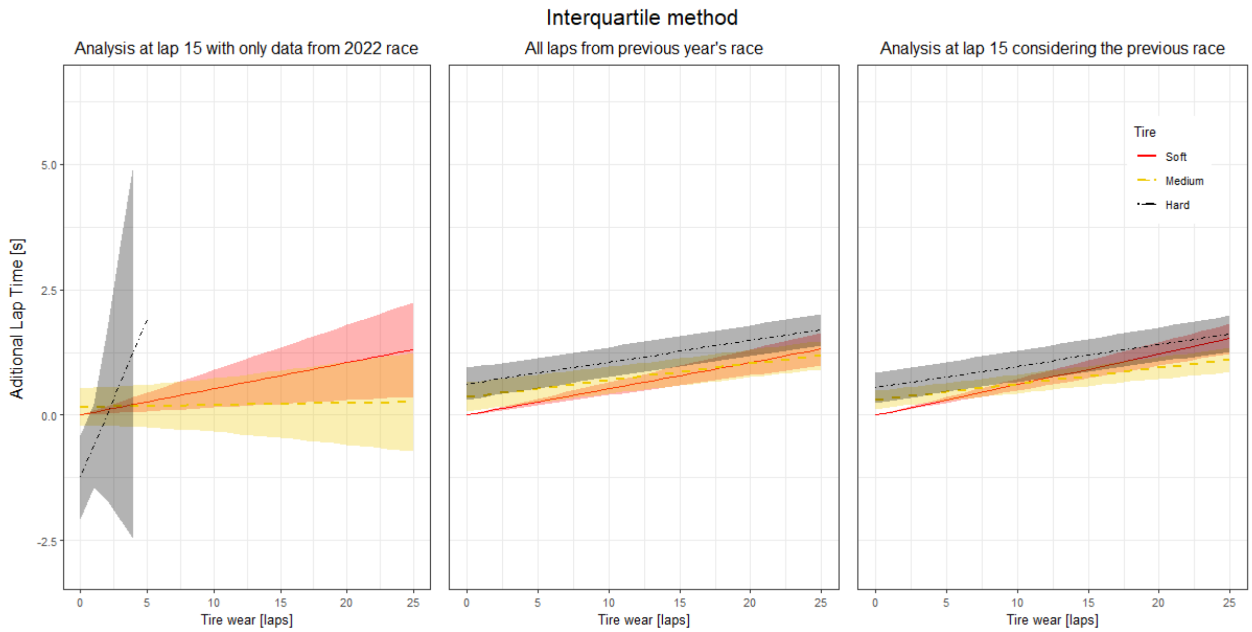

Figure C.2: Comparison of the tire performance at lap 15 Dutch Grand Prix 2022.

Continuing with the analysis, we can see the tire performance for all the tire types, at the end of the race, and considering data from the previous year's race.

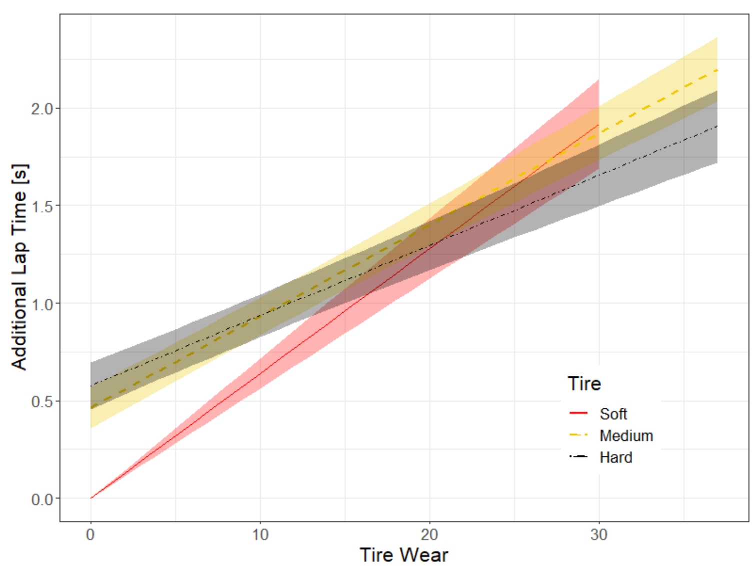

Figure C.3: Tire performance at the end of the Dutch Grand Prix 2022, considering data from the previous year's race

In the next Figure, we have the evolution of the race strategy for Max Verstappen for the Dutch Grand Prix 2022. Race that he ended up winning. As we can see, the yellow flags had a big impact on the tire strategy used. During the race, there were 2 yellow flags, and in both cases, the driver decided to make a pit stop. In the Figure we can see that from lap

25 the tire strategy converges to the use of a Soft tire at the beginning, then a Medium one, and to finish with a Soft tire compound. But, since there was a Yellow flag between lap 55 and 60, the team decided to put on a Hard tire compound to finish the race. And with the other Yellow Flag, they switched again to a Soft tire.

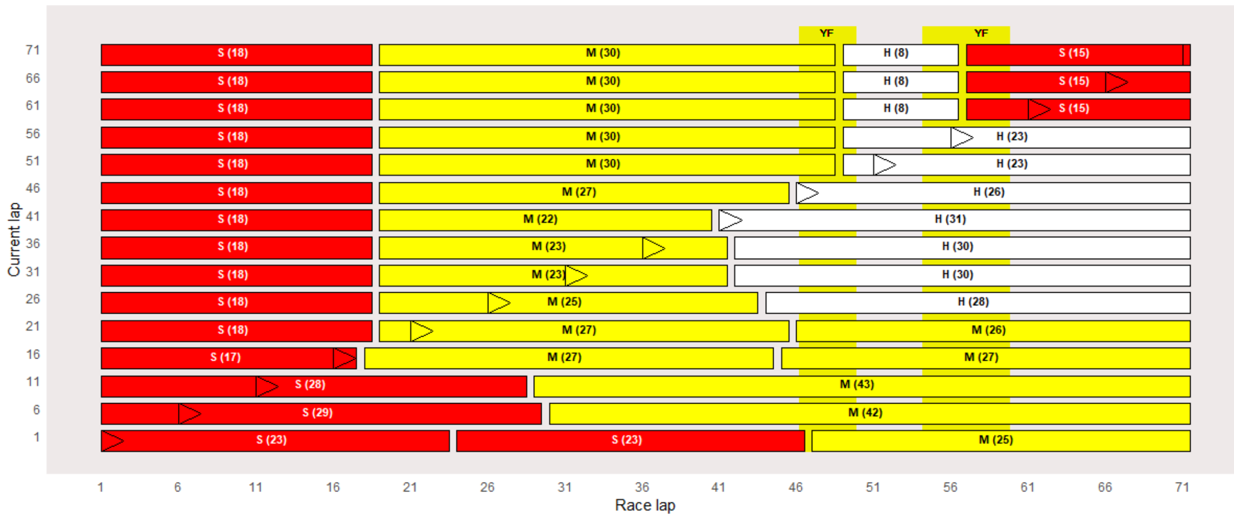

Figure C.4: Evolution of the race strategy for Max Verstappen for the Dutch Grand Prix 2022.

In Figure C.5 we show the comparison of the evolution of race strategy for Max Verstappen comparing the scenario where the prior race is taken into account and where it is not considered for the Dutch Grand Prix 2022.

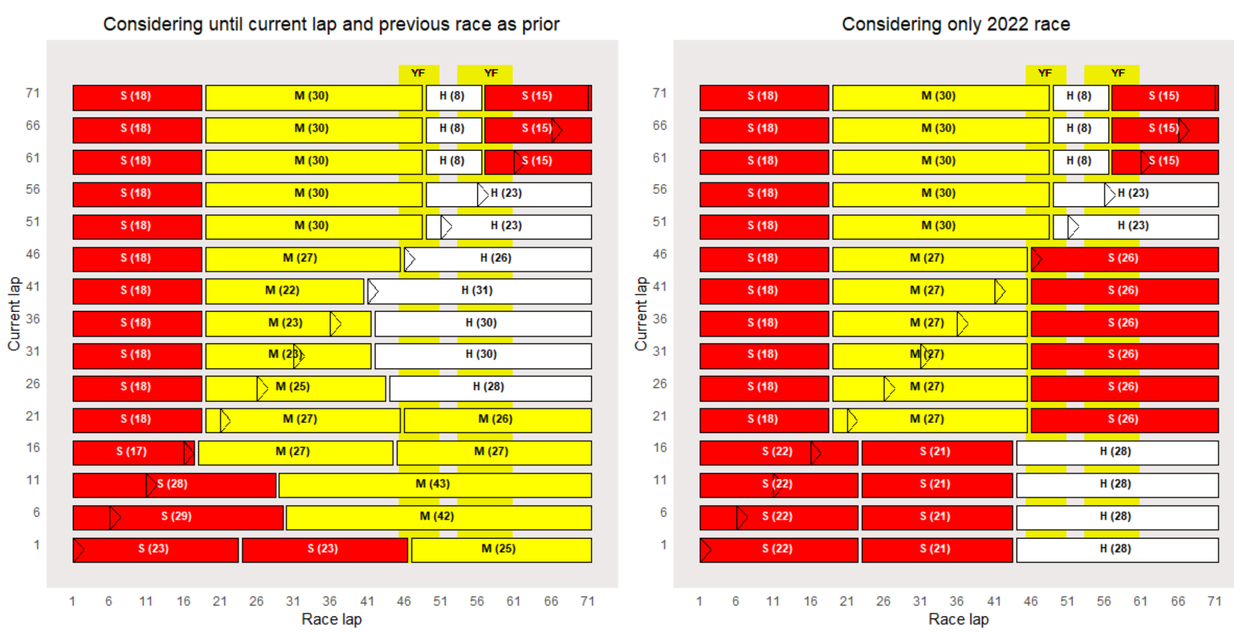

Figure C.5: Comparison of the evolution of race strategy for Max Verstappen comparing the scenario where the prior race is taken into account and where it is not considered for the Dutch Grand Prix 2022.

In the next Figure, we have the evolution of the race strategy for Lewis Hamilton for the Dutch Grand Prix 2022. In the Figure we can see that from lap 25 the tire strategy converges to the use of a Medium tire at the beginning, and then a Hard one. But, since there was a Yellow flag between lap 48 and 51, the team decided to put on a Medium tire compound to finish the race.

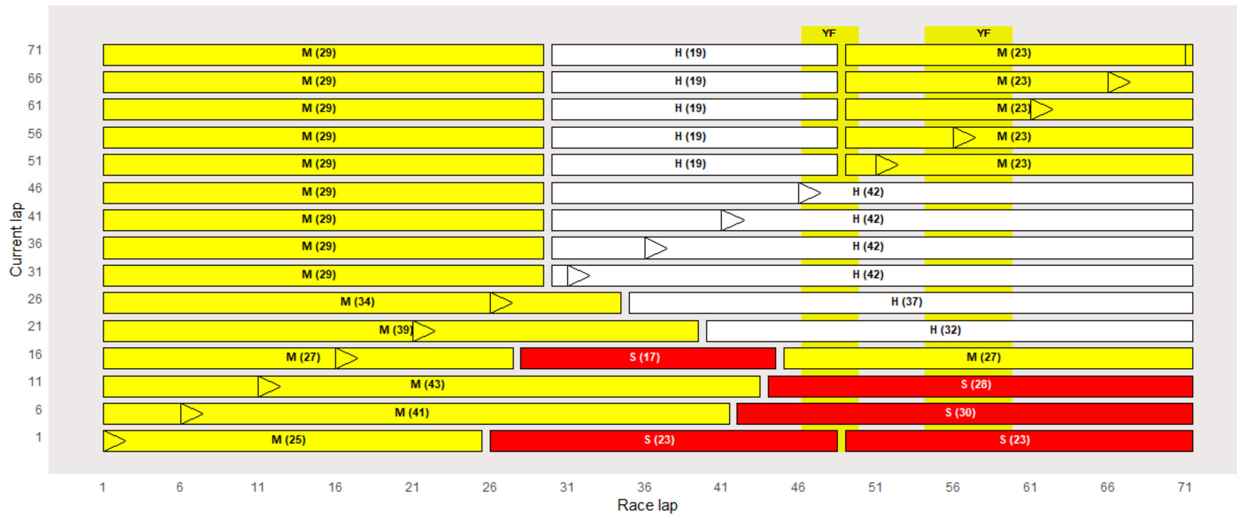

Figure C.6: Evolution of the race strategy for Lewis Hamilton for the Dutch Grand Prix 2022.

In Figure C.7 we show the comparison of the evolution of race strategy for Max Verstappen comparing the scenario where the prior race is taken into account and where it is not considered for the Dutch Grand Prix 2022.

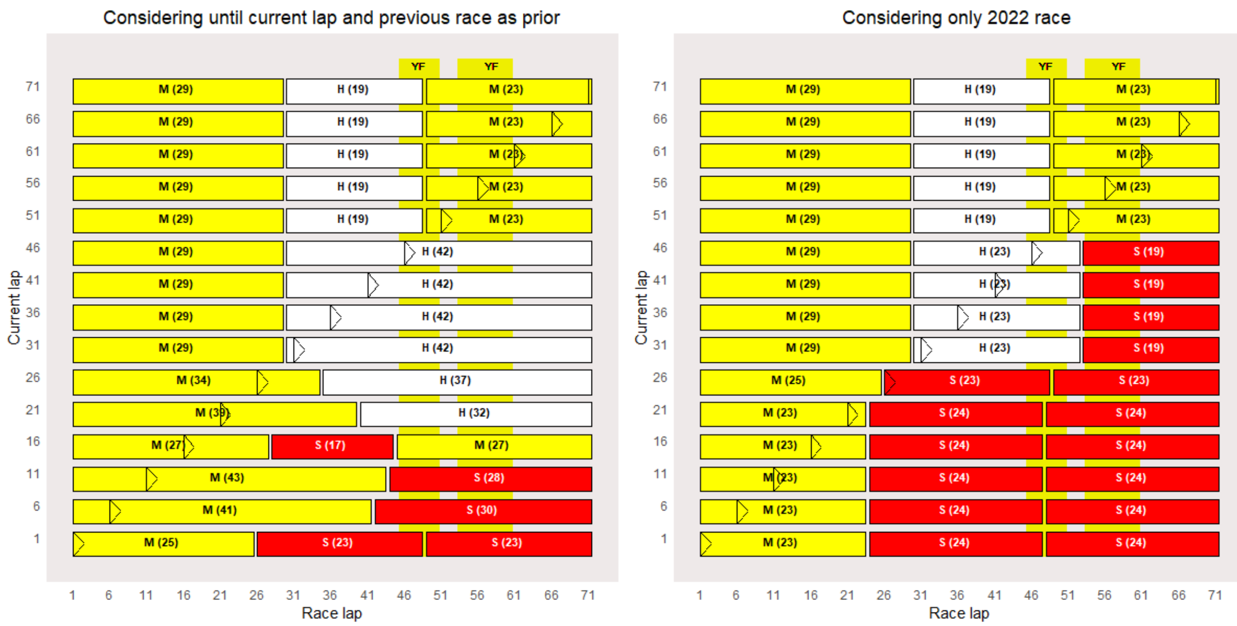

Figure C.7: Comparison of the evolution of race strategy for Lewis Hamilton comparing the scenario where the prior race is taken into account and where it is not considered for the Dutch Grand Prix 2022.

Continuing our analysis, we aim to evaluate how our online strategy compares to the optimal one for each driver of the grid. In Figure C.8, we illustrate the time difference between our online proposed strategy and the offline optimal strategy, with lap times calculated using the estimated coefficients from the offline case (i.e., the regression calibrated solely with 2022 lap time data without any prior information). It is noteworthy that our strategy becomes increasingly accurate as the race progresses, with a difference of no more than 5 seconds compared to the optimal one for most drivers. It is interesting to note that since lap 25, the strategy of the online setting aligns very closely with the offline strategy. Towards the end of the race, we observe that, for the majority of drivers, the time difference converges to zero.

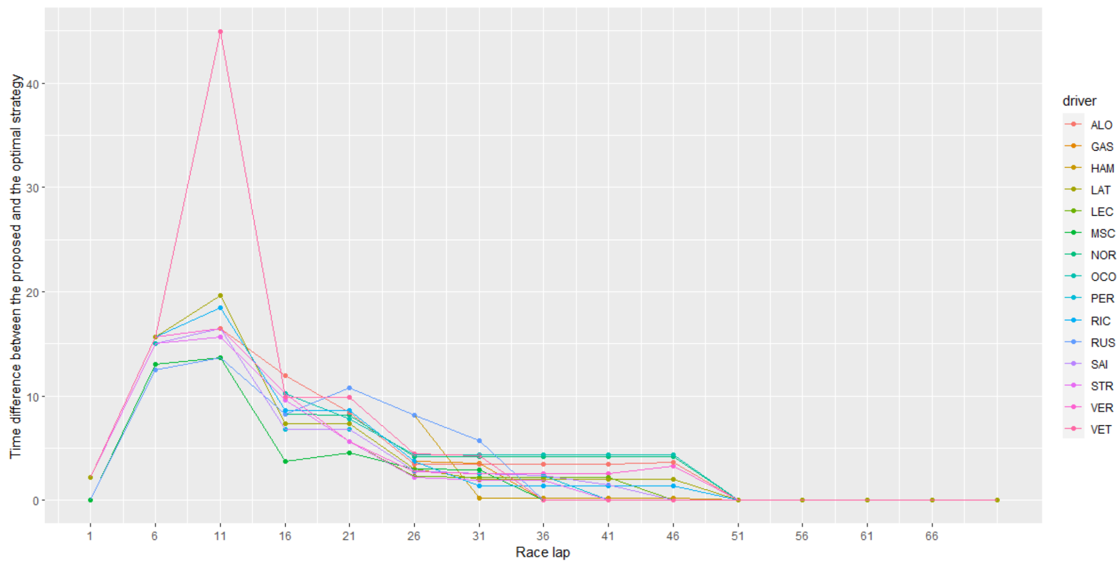

Figure C.8: Time difference between our proposed strategy and the optimal strategy during the Dutch Grand Prix 2022, evaluating every 5 laps.

In the figure below, on the left panel, we have the time difference between the real strategy used by the driver and the optimal strategy we propose (x-axis), for every driver (y-axis). On the right panel, we have the strategy used by each driver (y-axis) for the race (x-axis). Let us remember that the time difference is the additional time compared to using a new Soft tire compound in every lap. As we can see, on one hand, for drivers like Alonso, Ocon, Zhou and Stroll, our predicted strategy is very similar to the one used in the race. On the other hand, as mentioned before, we have Russell's race, a driver who didn't choose the best strategy. In fact, our strategy was more than 20 seconds faster than the real strategy.

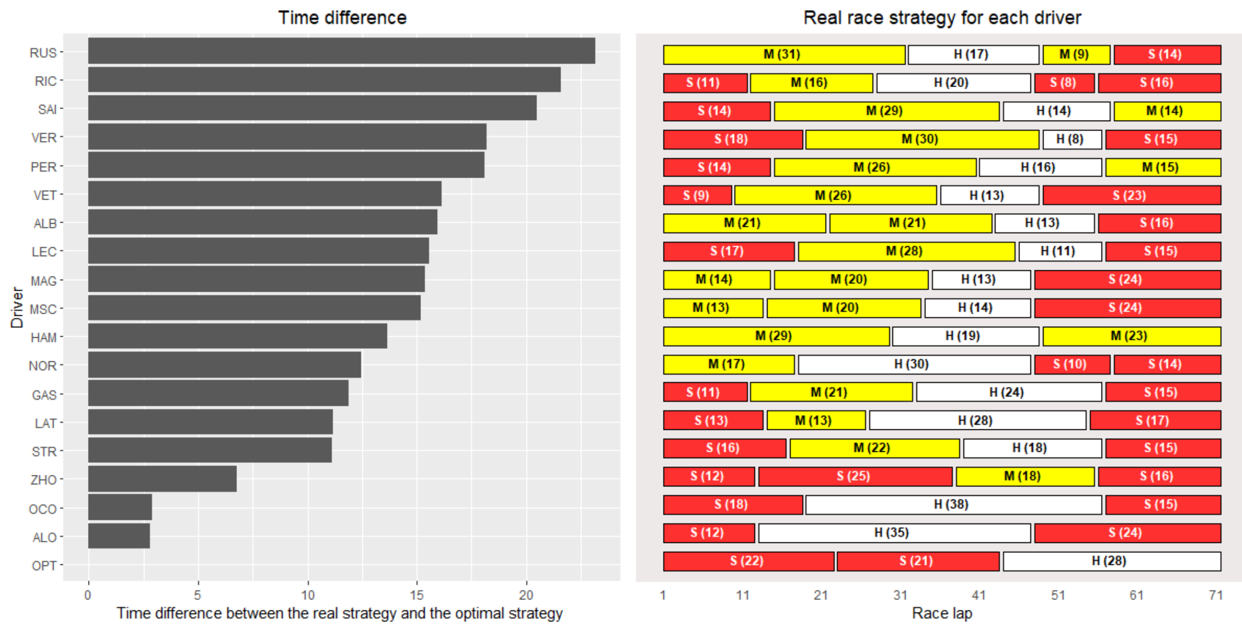

Figure C.9: Time difference between the real strategy used by the driver and the optimal strategy we propose for every driver during the Dutch Grand Prix 2022.

In Figure C.10, we show the time difference between the real strategy used by the driver and the optimal strategy we propose when using a regression with all the time laps of the Dutch GP of 2022 (x-axis), for every driver (y-axis). We are also showing on the x-axis the effect of the driver obtained from the regression. It is also important to notice that the drivers are ordered according to their finishing position during the Grand Prix. As we can see, there are some drivers, like Latifi, that even though they had a good strategy, the driver effect made their time increase. This is because even though a driver has a good strategy, their skills and the car can make a difference. In Figure C.11 we are also showing on the x-axis the effect of the team obtained from the regression. It is interesting to note that the effect of the driver makes a lot more sense than the team effect on the time difference.

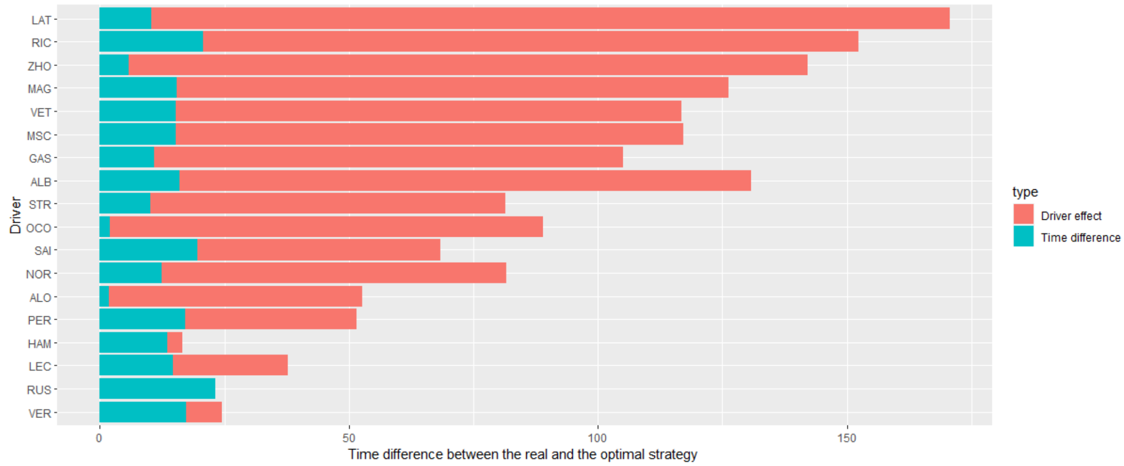

Figure C.10: Time difference between the real strategy used by the driver and the optimal strategy we propose for every driver during the Dutch Grand Prix 2022.

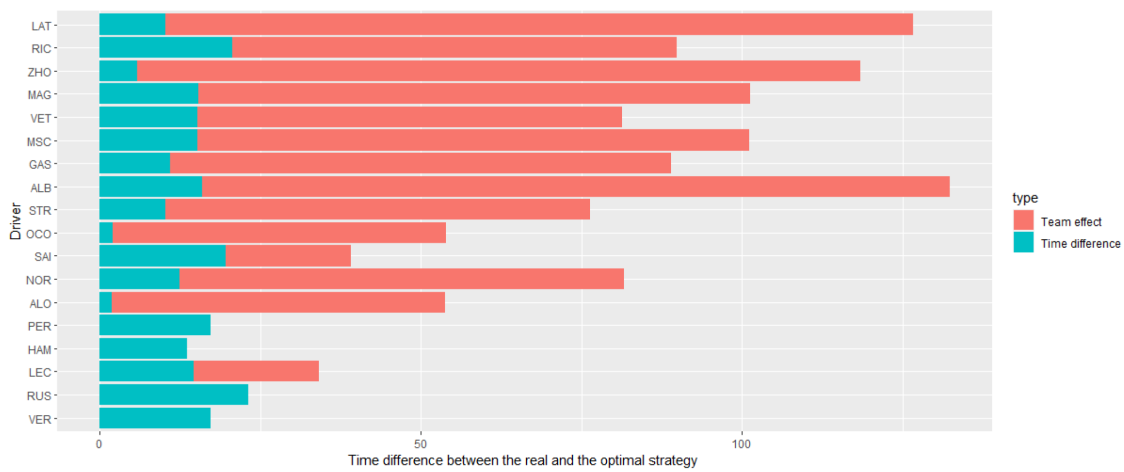

Figure C.11: Time difference between the real strategy used by the driver and the optimal strategy we propose for every driver during the Dutch Grand Prix 2022.