Universidade Federal de São Carlos
Centro de Ciências Exatas e de Tecnologia
Departamento de Estatística

UMA ABORDAGEM BASEADA EM PASSEIOS ALEATÓRIOS SOBRE GRAFOS PARA IDENTIFICAR O MELHOR PILOTO DE FÓRMULA 1 DA HISTÓRIA

Ricardo Teixeira Romanelli

Trabalho de Conclusão de Curso

Universidade Federal de São Carlos
Centro de Ciências Exatas e de Tecnologia
Departamento de Estatística

UMA ABORDAGEM BASEADA EM PASSEIOS ALEATÓRIOS SOBRE GRAFOS PARA IDENTIFICAR O MELHOR PILOTO DE FÓRMULA 1 DA HISTÓRIA

Ricardo Teixeira Romanelli
Orientador: Prof. Dr. Ricardo Felipe Ferreira

Trabalho de Conclusão de Curso apresentado ao Departamento de Estatística da Universidade Federal de São Carlos - DEs-UFSCar, como parte dos requisitos para obtenção do título de Bacharel em Estatística.

São Carlos
Setembro de 2024

Federal University of São Carlos
Exact and Technology Sciences Center
Department of Statistics

A RANDOM WALK-BASED APPROACH ON GRAPHS TO IDENTIFY THE BEST FORMULA 1 DRIVER OF ALL TIME.

Ricardo Teixeira Romanelli
Advisor: Prof. Ricardo Felipe Ferrerira

Bachelors dissertation submitted to the Department of Statistics, Federal University of São Carlos - DEs-UFSCar, in partial fulfillment of the requirements for the degree of Bachelor in Statistics.

São Carlos
September 2024

Ricardo Teixeira Romanelli

UMA ABORDAGEM BASEADA EM PASSEIOS ALEATÓRIOS SOBRE GRAFOS PARA IDENTIFICAR O MELHOR PILOTO DE FÓRMULA 1 DA HISTÓRIA

Este exemplar corresponde à redação final do trabalho de conclusão de curso devidamente corrigido e defendido por nome do(a) aluno(a) e aprovado pela banca examinadora.

Aprovado em 17 de setembro de 2024

Banca Examinadora:

- Nome do membro 1 Prof. Ricardo Felipe Ferreira
- Nome do membro 2 Prof. Alexsandro Giacomo Grimbert Gallo
- Nome do membro 3 Profa. Teresa Cristina Martins Dias

Dedico este trabalho aos meus pais, Ricardo e Luciene, e a minha irmã, Giulia, que foram fundamentais no meu percurso acadêmico e pessoal. Agradeço também ao meu grande amigo e companheiro, Gabriel, que esteve ao meu lado durante todos os anos acadêmicos me apoiando e me ajudando ao longo do curso.

Agradecimentos

Agradeço ao meu professor Ricardo, que me orientou com muita destreza e paciência ao longo do trabalho. Também agradeço aos dois professores componentes da Banca, Sandro, que me mostrou e ensinou a área de Processos Estocásticos, e Cris, que me proporcionou o ensino da melhor matéria da graduação, Probabilidade 2. Por último, agradeço ao piloto e meu ídolo Sebastian Vettel, que construiu minha paixão pelo esporte que deu início a este projeto.

"As pessoas inventam estatísticas para provar qualquer coisa.

40% das pessoas sabem disso!"

(Matt Groening)

Resumo

No mundo do automobilismo, quando estamos interessados em identificar o melhor piloto, levamos em consideração, em geral, somente aqueles que já correram na Fórmula 1, o ápice do esporte sobre quatro rodas mundial. No entanto, a definição do que significa ser o melhor piloto é subjetiva, pois pode envolver diferentes fatores sejam estes técnicos ou emocionais. Neste trabalho, propomos avaliar o desempenho dos pilotos de Fórmula 1 através de uma abordagem exata e objetiva. Para isso, a similaridade do desempenho de dois pilotos é mensurada a partir de medidas que levam em consideração variáveis que avaliam tanto o talento do piloto quanto a qualidade de seus carros. A partir dessas medidas de similaridade, vamos construir um grafo orientado com pesos em que cada vértice representa um piloto e cada aresta é ponderada por uma função da medida de similaridade entre os vértices que a define. Por fim, vamos encontrar a distribuição estacionária sobre esse grafo e o melhor piloto de Fórmula 1 é aquele que possui a maior distribuição. A partir dessa regra de decisão, obtemos também um *ranking* dos desempenhos dos pilotos.
Palavras-chave: *Fórmula 1; Grafos; Passeios aleatórios.*

Abstract

In the world of motorsports, when we are interested in identifying the best driver, we generally only consider those who have raced in Formula 1, the pinnacle of the global four-wheeled sport. However, the definition of what it means to be the best driver is subjective, as it can involve different factors, whether technical or emotional. In this work, we propose to evaluate the performance of Formula 1 drivers through an exact and objective approach. To do this, the similarity of the performance of two drivers is measured based on metrics that take into account variables that evaluate both the driver's talent and the quality of their cars. From these similarity measures, we will build a directed graph with weights, where each vertex represents a driver and each edge is weighted by a function of the similarity measure between the vertices that defines it. Finally, we will find the stationary distribution over this graph, and the best Formula 1 driver is the one with the highest distribution. Based on this decision rule, we also obtain a ranking of driver performance.

Keywords: *Formula 1; Graphs; Random walks.*

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

A Fórmula 1 é o campeonato automobilístico mais veloz do mundo (Piquero *et al.*, 2021). É a categoria mais avançada do esporte a motor e é regulamentada pela Federação Internacional de Automobilismo (FIA). O Campeonato Mundial de Fórmula 1 da FIA tem sido uma das principais categorias de corrida em todo o mundo e um dos mais assistidos de todos os esportes televisionados (Piquero *et al.*, 2021).

Uma temporada da Fórmula 1 consiste em uma série de corridas, conhecidas como Grandes Prêmios, que acontecem ao redor do mundo em circuitos construídos para esse fim, ou em ruas fechadas. Devido a este apelo internacional, não é surpresa que os fãs e a própria mídia, especializada no esporte, gostem de classificar as equipes e os pilotos mais bem sucedidos de todos os tempos, dentre todos aqueles que já competiram na história. No entanto, existem poucos estudos empíricos que desenvolveram ou aplicaram técnicas estatísticas rigorosas para analisar quais os corredores de Fórmula 1 foram os mais bem sucedidos (Eichenberger *et al.*, 2009; Phillips, 2014; Bell *et al.*, 2016; Piquero *et al.*, 2021).

A dificuldade em identificar o melhor piloto de Fórmula 1 da história está atrelada ao fato de que o desempenho observável de um piloto depende tanto do seu talento quanto de outras variáveis externas, tais como a qualidade de seus carros e a competitividade de seus rivais. Além disso, existe uma dificuldade inerente na definição do que significa ser o melhor. Assim, as listas e metodologias utilizadas para identificar o melhor piloto são variadas e, algumas vezes, controversas. Existe também a dificuldade natural em comparar atletas de temporadas em épocas distintas, uma vez que as próprias regras, regulamento e maquinário utilizado sofreram e ainda sofrem mudanças ao longo do tempo. Uma forma de superar essa dificuldade é considerar anos comparáveis, para manter o maior número de regras e regulamentos iguais ou o mais padronizado possível. Neste trabalho, usamos

uma metodologia baseada em grafos para levar em consideração o efeito temporal que existe nos campeonatos sobre a performance dos pilotos.

Cada piloto considerado em nossa análise estatística representará um vértice de um grafo. Se dois pilotos já correram juntos em uma mesma temporada, então existirá uma aresta conectando estes dois vértices. Assim, dois pilotos estarão relacionados se, e somente se, tiverem sido oponentes em alguma temporada. A força dessa relação é definida a partir de uma medida de similaridade que, com base em um conjunto de covariáveis, informa o quão próxima é a performance dos pares de pilotos. Note que, dessa forma, estamos levando em consideração o efeito temporal das temporadas sobre a performance dos pilotos, uma vez que estamos mensurando a similaridade apenas entre pilotos que competiram juntos em uma ou mais temporadas. A regra que utilizamos para decidir qual é o melhor piloto da história é baseada em passeios aleatórios por meio de uma matriz cujas probabilidades de transição entre dois vértices são funções das medidas de similaridade entre estes. Cada passeio aleatório iniciará sua trajetória em um vértice diferente do grafo e, após uma quantidade suficientemente grande de transições, analisamos quais foram os vértices mais visitados e, a partir deste número de visitas, identificamos quais são os melhores pilotos de todos os tempos.

O conjunto de dados que utilizamos neste trabalho pode ser encontrado no repositório Kaggle, disponível em Kaggle - Formula 1 Data Base. Esse conjunto contém informações de 854 pilotos e é composto por diversas variáveis contemplando as temporadas de 1950 até 2022. Neste trabalho, enfrentamos o problema de selecionar ou construir as variáveis que utilizamos para definir a medida de similaridade, levando em consideração a opinião de *sites* especializados no assunto e, possivelmente, de especialistas que trabalham com o esporte e estão familiarizados com esse tipo de conjunto de dados.

A principal contribuição deste trabalho é o uso de uma metodologia baseada em passeios aleatórios sobre grafos para identificar o melhor corredor da história. Após pesquisar, não foi encontrado trabalho na literatura que utilize essa teoria com tal finalidade. Nesse sentido, esse trabalho complementa os estudos sobre métodos estatísticos aplicados aos esportes.

Este trabalho está organizado da seguinte maneira. No próximo capítulo, apresentamos uma breve história da Fórmula 1, das equipes participantes do esporte e as modificações sofrida pelo sistema de pontuação ao longo do tempo. Propomos também, uma forma de padronizar o sistema de pontuação para tornar mais justa a comparação das per-

formances dos pilotos. No Capítulo 3, apresentamos a metodologia baseada em simulação de passeios aleatórios sobre grafos, que é utilizada para ranquear os pilotos de Fórmula 1 de acordo com sua performance. No Capítulo 4 mostramos um passo a passo do ranqueamento desenvolvido com uma amostra menor (5 pilotos), para melhor explicação do desenvolvimento. No Capítulo 5, foi desenvolvido o ranqueamento partindo da amostra de todos os pilotos, junto com uma análise de sensibilidade de acordo com certas mudanças feitas ao longo do processo. Por último, no Capítulo 6, concluímos o trabalho discutindo os resultados e propondo próximos passos para uma eventual continuação da pesquisa.

Capítulo 2

Fórmula 1

Para o melhor entendimento do processo de ranqueamento dos pilotos que foi desenvolvido neste estudo, é importante entendermos como a Fórmula 1 funciona, desde sua história, o conjunto de regras do campeonato até quais são os pilotos, equipes e fabricantes que ganharam títulos e se consolidaram como pessoas marcantes na história do esporte. Nesse sentido, foram selecionadas características dos pilotos que foram levadas em consideração durante a aplicação da metodologia proposta neste estudo. A relevância de tais características para o estudo dar-se-á de acordo com o que, a partir da literatura vigente, consideramos que define o que é ser o melhor piloto.

2.1 História da F1

O campeonato de Fórmula 1 começou a ser disputado em 1950, quando a Federação Internacional de Automobilismo (FIA) decidiu reunir uma série de grandes prêmios (GPs) em um único torneio mundial. GPs são formados por uma série de eventos que ocorrem ao longo de um fim de semana em um mesmo autódromo, com o principal evento sendo a corrida. "Fórmula" é o nome dado para representar o conjunto de regras que todas as fabricantes devem seguir para poder participar da competição, visando uma competição equilibrada.

O primeiro campeonato de Fórmula 1 ocorreu em 1950, com 7 GPs. O primeiro evento ocorreu no circuito de Silverstone em um fim de semana de Maio e o vencedor foi o piloto Nino Farina, pilotando um Alfa Romeo. Outras equipes/fabricantes que participavam na época (algumas participam até hoje) eram Ferrari, Maserati e Mercedes. Ao longo da década de 50, o campeonato começou a ser disputado em países fora da

Europa, aumentando o número de GPs em um mesmo ano. Países como Estados Unidos, Argentina e Marrocos foram adicionados ao calendário ao longo dos anos e, em 2022, tivemos um total de 22 GPs disputados em 4 continentes, o maior número de provas em um único ano até então.

É importante ressaltar a diferença da quantidade de corridas em uma temporada, pois ao comparar estatísticas entre um piloto de antigamente com um piloto de anos mais atuais, ambos podem ter corrido por um mesmo período de tempo, porém participado de um número totalmente diferente de corridas, o que faz com que as estatísticas absolutas enganem. Um exemplo claro disso é que, por mais que o argentino Juan Manuel Fangio, corredor da década de 50, tenha 24 vitórias e o inglês Lewis Hamilton tenha 103 vitórias, o percentual de vitórias de Fangio é maior que o percentual de vitórias de Hamilton ($47,05\%$ e $33,22\%$, respectivamente). Ao longo do trabalho, métodos foram desenvolvidos para que possamos compará-los sem enfrentar problemas, como a diferença absoluta do número de corridas participadas.

Para entender um pouco melhor a evolução da Fórmula 1 ao longo dos anos, é apresentado uma ideia geral do panorama da competição para cada década, as maiores evoluções tecnológicas que mudaram o "status quo" das corridas e os pilotos que tiveram destaque, tanto na época que corriam, quanto atualmente em publicações que tentam ranquear os melhores pilotos da história.

2.1.1 Anos 50

A década de 1950 foi marcada pela expansão inicial da competição, na qual o número de corridas foi de 7 para 11 em um curto período de tempo. A maior inovação da época em quesitos técnicos foi a troca do lugar do motor, que antes ficava na parte dianteira, para a parte traseira do carro. Tal inovação foi feita pela equipe Cooper, pilotada pelo australiano Jack Brabham. Outros pilotos de expressão da época foram Nino Farina, já mencionado anteriormente, Juan Manuel Fangio, um dos maiores campeões até hoje com 5 títulos, Mike Hawthorn, Alberto Ascari e Bruce McLaren, que futuramente teria sua própria equipe.

Pilotos Marcantes da Época

- Juan Manuel Fangio (1950 - 1958).

  - Títulos: 5.
  - Corridas: 51.
  - Vitórias: 24 (47,05%).
  - Pole positions: 29 (56,86%).
  - Pódios: 35 (68,62%).
  - Pontos na carreira: 277,6.
  - Voltas mais rápidas: 23 (45,09%).

- Alberto Ascari (1950 - 1955).
  - Títulos: 2.
  - Corridas: 33.
  - Vitórias: 13 (39,39%).
  - Pole positions: 14 (42,42%).
  - Pódios: 17 (51,52%).
  - Pontos na carreira: 140,1.
  - Voltas mais rápidas: 12 (36,36%).

- Jack Brabham (1955 - 1970).
  - Títulos: 3.
  - Corridas: 128.
  - Vitórias: 14 (10,93%).
  - Pole positions: 13 (10,15%).
  - Pódios: 31 (24,21%).
  - Pontos na carreira: 253.
  - Voltas mais rápidas: 12 (9,37%).

2.1.2 Anos 60

O *site* Formula One Art & Genius (Manishin, 2018b) define os anos 60 como "a era britânica" devido a dominância pelos pilotos ingleses, tais como Graham Hill, Jim Clark,

John Surtees e Jackie Stewart pilotando carros de equipes também inglesas como Cooper, BRM, Brabham e Lotus. Juntos, tais pilotos colecionaram 6 títulos entre 1961 e 1970. As inovações técnicas de maior destaque na época foram a mudança do banco no qual os pilotos sentam, chamado de *cockpit* e a criação do aerofólio traseiro. Agora, os bancos são mais inclinados e confortáveis, diferente dos bancos da década passada que eram retos, na posição de 90 graus. Além disso, pequenas peças de metal no formato de asa foram adicionados na parte de trás do carro para que ajudasse o carro a ter uma melhor aerodinâmica, deixando-o mais rápido.

Pilotos Marcantes da Época

- Jim Clark (1960 - 1968).
  - Títulos: 2.
  - Corridas: 73.
  - Vitórias: 25 (34,24%).
  - Pole positions: 33 (45,20%).
  - Pódios: 32 (43,83%).
  - Pontos na carreira: 255.
  - Voltas mais rápidas: 10 (13,69%).
- Graham Hill (1958 - 1975).
  - Títulos: 2.
  - Corridas: 179.
  - Vitórias: 14 (7,82%).
  - Pole positions: 13 (7,26%).
  - Pódios: 36 (20,11%).
  - Pontos na carreira: 270.
  - Voltas mais rápidas: 10 (5,58%).
- Jackie Stewart (1965 - 1973).
  - Títulos: 3.

  - Corridas: 99.
  - Vitórias: 27 (27,27%).
  - Pole positions: 17 (17,17%).
  - Pódios: 43 (43,43%).
  - Pontos na carreira: 360.
  - Voltas mais rápidas: 15 (15,15%).

2.1.3 Anos 70

Nos anos 70, o Brasil sediou seu primeiro GP, um marco importante não só para o país, mas para a Fórmula 1 como um todo, dado que o autódromo José Carlos Pace (ou Interlagos), pista onde o país sedia seu GP, é considerado por muitos como um dos melhores já existentes. O colunista Nathan Reynolds, editor do *site* Last Word on Sports (Reynolds, 2022), construiu um ranking de melhores pistas para a Fórmula 1 e Interlagos está em primeiro. Além do autódromo estreando na competição, o primeiro brasileiro de destaque no mundo automobilístico, Emerson Fittipaldi, foi bicampeão mundial: a primeira vez pela equipe Lotus e a segunda pela equipe McLaren. Outros nomes que fizeram história foram Niki Lauda e James Hunt, que, de acordo com o jornal Express UK (Leathers, 2022), "possuem a rivalidade mais conhecida na história da Fórmula 1". Mario Andretti, Gilles Villeneuve e Jody Sheckter foram outros nomes que dominaram o topo das tabelas na época. Ao longo da história do esporte, observamos várias dominâncias e dinastias serem construídas. Uma delas foi a da fabricante Ford, que, com seu motor, ganhou 12 títulos entre 1968 e 1982.

Pilotos Marcantes da Época

- Emerson Fittipaldi (1970 - 1980).
  - Títulos: 2.
  - Corridas: 149.
  - Vitórias: 14 (9,39%).
  - Pole positions: 6 (4,02%).
  - Pódios: 35 (23,48%).

  - Pontos na carreira: 281.
  - Voltas mais rápidas: 6 (4,02%).

- Niki Lauda (1971 - 1979 / 1892 - 1985).
  - Títulos: 3.
  - Corridas: 177.
  - Vitórias: 25 (14,12%).
  - Pole positions: 24 (13,55%).
  - Pódios: 52 (29,37%).
  - Pontos na carreira: 420.
  - Voltas mais rápidas: 24 (13,55%).

- James Hunt (1973 - 1979).
  - Títulos: 1.
  - Corridas: 93.
  - Vitórias: 10 (10,75%).
  - Pole positions: 14 (15,05%).
  - Pódios: 23 (24,73%).
  - Pontos na carreira: 179.
  - Voltas mais rápidas: 8 (8,60%).

2.1.4 Anos 80

Os anos 80 foi conquistado na sua grande maioria por brasileiros e franceses na competição, em especial, os brasileiros Nelson Piquet e Ayrton Senna e o francês Alain Prost que, juntos, colecionaram 7 títulos na década, 2 para cada brasileiro e 3 para o francês (Senna e Prost ganham mais um título cada nos anos 90). A equipe a ser batida agora era a McLaren-Honda. A junção entre a equipe inglesa e a fabricante de motor japonesa ganhou 5 títulos seguidos ao final da década. De acordo com a matéria produzida para o Globo Esporte (Lopes, 2022) por Rafael Lopes, comentarista de automobilismo do Grupo Globo, Senna e Prost protagonizaram a maior rivalidade da história do esporte mundial.

Tal dominância e rivalidade foi coroada em 1988 com o que muitos, como o *site* Bleacher Report (Harden, 2015), consideram a maior dominância de uma equipe e pilotos na história das corridas de grandes prêmios, na qual a equipe Mclaren-Honda ganhou 15 das 16 corridas da edição do campeonato, com Senna ganhando 8 corridas e Prost ganhando 7. Na corrida não ganhada pela McLaren, ambos os carros abandonaram a corrida antes de seu término.

Pilotos Marcantes da Época

- Nelson Piquet (1978 - 1991).
  - Títulos: 3.
  - Corridas: 207.
  - Vitórias: 23 (11,11%).
  - Pole positions: 24 (11,59%).
  - Pódios: 60 (28,98%).
  - Pontos na carreira: 485,5.
  - Voltas mais rápidas: 23 (11,11%).
- Alain Prost (1980 - 1991 / 1993).
  - Títulos: 4.
  - Corridas: 202.
  - Vitórias: 51 (25,24%).
  - Pole positions: 33 (15,94%).
  - Pódios: 106 (51,20%).
  - Pontos na carreira: 798.
  - Voltas mais rápidas: 41 (19,80%).
- Ayrton Senna (1984 - 1994).
  - Títulos: 3.
  - Corridas: 162.

  - Vitórias: 41 (25,30%).
  - Pole positions: 65 (40,12%).
  - Pódios: 80 (49,38%).
  - Pontos na carreira: 610.
  - Voltas mais rápidas: 19 (11,72%).

2.1.5 Anos 90

Os anos 90 foram marcados por muitas inovações tecnológicas, como a suspensão ativa, que ajudou a equipe Williams quebrar a dinastia McLaren-Honda com os pilotos Nigel Mansell e Alain Prost, que agora corria não mais pela McLaren e sim pela Williams, em seu último título. Tais inovações deixaram os carros complexos e rápidos, porém, com isso, se tornaram também mais perigoso. O resultado disso foi um esporte cada vez mais imprevisível e perigoso. A página Formula One Art & Genius (Manishin, 2018a) diz que, após o primeiro fim de semana de maio de 1994, as mortes de Roland Ratzenberger e Ayrton Senna culminaram em uma revolução na segurança dos carros, diminuindo drasticamente os acidentes graves que eram de costume. No fim da década, o alemão Michael Schumacher, primeiramente com a equipe Benneton e depois com a equipe italiana Ferrari deu início ao período de nova dominância de um piloto e ao fim da seca de títulos de 21 anos da equipe italiana.

Pilotos Marcantes da Época

- Nigel Mansell (1980 - 1992 / 1994 - 1995).
  - Títulos: 1.
  - Corridas: 192.
  - Vitórias: 31 (16,14%).
  - Pole positions: 32 (16,67%).
  - Pódios: 59 (36,41%).
  - Pontos na carreira: 482.
  - Voltas mais rápidas: 30 (18,51%).
- Mika Hakkinen (1991 - 2001).

  - Títulos: 2.
  - Corridas: 165.
  - Vitórias: 20 (12,12%).
  - Pole positions: 26 (15,75%).
  - Pódios: 51 (30,90%).
  - Pontos na carreira: 420.
  - Voltas mais rápidas: 25 (15,15%).

- Jacques Villeneuve (1996 - 2006).
  - Títulos: 1.
  - Corridas: 165.
  - Vitórias: 11 (6,67%).
  - Pole positions: 12 (7,27%).
  - Pódios: 23 (14,19%).
  - Pontos na carreira: 235.
  - Voltas mais rápidas: 9 (5,56%).

2.1.6 Anos 2000

Nos anos 2000, a Scuderia Ferrari dominou o início da década, ganhando os títulos de 2000 até 2004 com Schumacher. Após o tempo definido por Edd Straw, colunista do *site* AutoSport (Straw, 2017), como "Supremacia Ferrari", novas ascenções de pilotos como Fernando Alonso, Lewis Hamilton e Sebastian Vettel tomaram presença, em um curto período em que o equilíbrio das equipes prevaleciam, com 5 equipes e 5 pilotos diferentes sendo campeões entre 2006 e 2010.

Pilotos Marcantes da Época

- Michael Schumacher (1991 - 2006 / 2010 - 2012).
  - Títulos: 7.
  - Corridas: 308.

  - Vitórias: 91 (29,54%).
  - Pole positions: 68 (22,07%).
  - Pódios: 155 (50,32%).
  - Pontos na carreira: 1566.
  - Voltas mais rápidas: 77 (25%).

- Kimi Raikkonen (2001 - 2009 / 2012 - 2021).
  - Títulos: 1.
  - Corridas: 342.
  - Vitórias: 21 (6,41%).
  - Pole positions: 18 (5,82%).
  - Pódios: 103 (30,11%).
  - Pontos na carreira: 1863.
  - Voltas mais rápidas: 46 (13,45%).

- Fernando Alonso (2001 / 2003 - 2018 / 2021 - Atualmente).
  - Títulos: 2.
  - Corridas: 359.
  - Vitórias: 32 (8,91%).
  - Pole positions: 22 (6,12%).
  - Pódios: 98 (27,29%).
  - Pontos na carreira: 2061.
  - Voltas mais rápidas: 23 (6,40%).

2.1.7 Anos 2010 - Atualidade

Os Anos 2010 foram marcados pela dominância da equipe Red Bull Racing junto com o alemão Sebastian Vettel ganhando 4 títulos entre 2010 e 2013 e a criação do carro híbrido. Nesta década, três motores foram utilizados pelos carros, um movido a gasolina e dois elétricos. Junto com a era dos carros híbridos iniciada em 2014, a equipe

Mercedes, que voltava de um grande hiato sem participar dos campeonatos, ganhou todos os campeonatos até 2021, sua grande maioria com Hamilton, que se tornou, empatado com Schumacher, o maior detentor de títulos da história, sendo heptacampeão. Em 2021, essa nova dinastia foi quebrada com a Red Bull e com o maior piloto em ascensão hoje em dia, Max Verstappen.

Pilotos Marcantes da Época

- Sebastian Vettel (2007 - 2022).
  - Títulos: 4.
  - Corridas: 300.
  - Vitórias: 53 (17,67%).
  - Pole positions: 57 (19%).
  - Pódios: 122 (40,67%).
  - Pontos na carreira: 3098.
  - Voltas mais rápidas: 38 (12,67%).
- Lewis Hamilton (2007 - Atualmente).
  - Títulos: 7.
  - Corridas: 310.
  - Vitórias: 103 (33,22%).
  - Pole positions: 103 (33,22%).
  - Pódios: 191 (61,61%).
  - Pontos na carreira: 4405.
  - Voltas mais rápidas: 61 (19,67%).
- Max Verstappen (2015 - Atualmente).
  - Títulos: 2.
  - Corridas: 163.
  - Vitórias: 35 (21,47%).

– Pole positions: 20 (12,26%).

– Pódios: 77 (47,23%).

– Pontos na carreira: 2011,5.

– Voltas mais rápidas: 21 (12,88%).

2.2 Equipes

Na Fórmula 1, para que um piloto alcance a glória e seja campeão, é necessário além de talento, um bom carro. Uma estatística que mostra isso é que dos 73 títulos disputados, em 84,93% das vezes, o piloto campeão era da equipe que também acabou se tornando campeã no ano, ou seja, apenas 11 vezes um piloto foi campeão em uma equipe que não tenha ganhado o título de construtores também.

Como discutido anteriormente, a evolução tecnológica anda lado a lado com a competição desde seus primórdios. Como consequência, um carro de décadas atrás não conseguiria competir com um carro atual, tanto por conta da mudança dos regulamentos quanto por conta dos avanços tecnológicos que fazem os carros de hoje em dia alcançarem velocidade que, até pouco tempo atrás, eram inimagináveis. Um exemplo disso é o gráfico onde é possível notar as melhores voltas feitas em cada corrida que ocorreu em Interlagos.

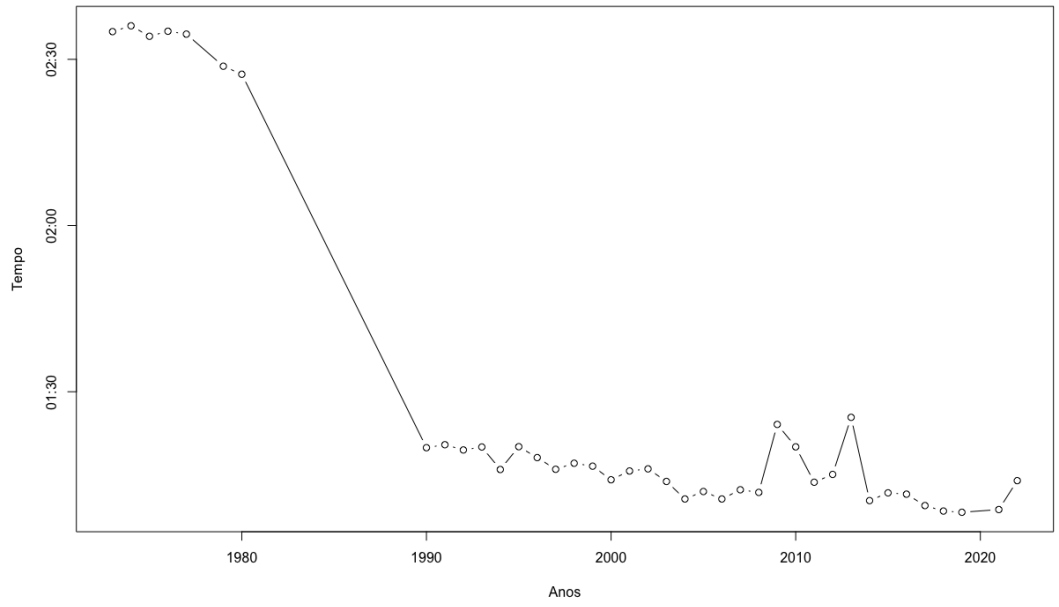

Figura 2.1: Comportamento dos tempos das melhores voltas por ano dos ganhadores em Interlagos mensurado em minutos.

Como podemos observar, o tempo da melhor volta no autódromo brasileiro de um carro dos anos 80 é mais de um minuto mais devagar que a de um carro atual, o que pode não parecer muito, mas é quase o dobro do tempo de volta do carro atual. Até mesmo olhando para os últimos 20 anos, percebemos uma queda circunstancial nos tempos de volta, mostrando que a cada ano os engenheiros estão procurando sempre evoluir seus carros, buscando a perfeição aerodinâmica, não sendo justo comparar pilotos ou equipes de décadas ou anos diferentes.

Equipes de destaque na história

- Scuderia Ferrari.
  - Títulos: 15.
  - Corridas: 1054.
  - Vitórias: 241 (22,86%).
  - Pole positions: 242 (22,96%).
  - Pódios: 793 (75,23%).
  - Voltas mais rápidas: 258 (24,47%).
- McLaren Racing.
  - Títulos: 8.
  - Corridas: 928.
  - Vitórias: 183 (19,71%).
  - Pole positions: 156 (16,81%).
  - Pódios: 494 (53,23%).
  - Voltas mais rápidas: 162 (17,45%).
- Mercedes AMG F1 Team.
  - Títulos: 8.
  - Corridas: 271.
  - Vitórias: 125 (46,12%).
  - Pole positions: 136 (50,18%).

  - Pódios: 281 (103,69%).
  - Voltas mais rápidas: 100 (36,90%).

- Williams Racing.
  - Títulos: 7.
  - Corridas: 792.
  - Vitórias: 114 (14,39%).
  - Pole positions: 128 (16,16%).
  - Pódios: 313 (39,52%).
  - Voltas mais rápidas: 133 (16,79%).

- Team Lotus.
  - Títulos: 7.
  - Corridas: 491.
  - Vitórias: 79 (16,08%).
  - Pole positions: 107 (21,79%).
  - Pódios: 200 (40,73%).
  - Voltas mais rápidas: 71 (14,46%).

- Red Bull Racing.
  - Títulos: 5.
  - Corridas: 348.
  - Vitórias: 92 (26,43%).
  - Pole positions: 81 (23,27%).
  - Pódios: 234 (67,24%).
  - Voltas mais rápidas: 84 (24,13%).

2.3 Sistema de Pontuação e Correção de Variáveis

O sistema de pontuação na competição mudou ao longo do tempo, de acordo com as alterações de regulamento que ocorriam. Segue a Tabela 2.3 sendo possível observar como as pontuações eram dadas desde 1950 até os dias de hoje, na qual a pontuação segue uma padronização criada pela FIA em 2010.

Tabela 2.1: Sistema de pontuação ao longo dos anos.

	1º	2º	3º	4º	5º	6º	7º	8º	9º	10º
1950-1959	8	6	4	3	2	-	-	-	-	-
1960	8	6	4	3	2	1	-	-	-	-
1961-1990	9	6	4	3	2	1	-	-	-	-
1991-2002	10	6	4	3	2	1	-	-	-	-
2003-2009	10	8	6	5	4	3	2	1	-	-
2010-Atual	25	18	15	12	10	8	6	4	2	1

Tais mudanças encontradas na história podem ser um problema na hora da comparação entre dois pilotos, principalmente ao comparar pilotos pré 2010 com pilotos pós 2010. Por exemplo, ao comparar o holandês Max Verstappen e o francês Alain Prost, mesmo Prost tendo o dobro de títulos e um maior número de corridas que o holandês, por conta da padronização para pontuações mais altas recentemente, Verstappen possui mais que o dobro de pontos que Prost (2011,5 e 798, respectivamente). Um piloto pós 2010 pode, de fato, possuir uma carreira melhor que um piloto pré 2010 e ter uma pontuação mais elevada, porém o certo é compará-los seguindo o mesmo sistema de pontuação.

2.3.1 Padronização do sistema de pontuação

Dado que o sistema de pontuação atual é o único que pontuam aqueles que chegaram até o décimo lugar, é de bom tamanho utilizá-lo para que um maior número de informações seja fornecido ao trabalhar com tal variável. Para pilotos mais antigos, que pontuavam muito menos, a diferença de pontuação com e sem a padronização se torna considerável. A revista Super Interessante (Rossini, 2023), em uma reportagem escrita por Maria Clara Rossini, mostra que Ayrton Senna, por exemplo, obteve 614 pontos ao longo de sua carreira. Caso ele tivesse pontuado conforme a pontuação atual, teria 1874 pontos, o triplo do que ele realmente pontuou ao longo de suas corridas.

O curioso ao fazer essa mudança de pontuação nos campeonatos anteriores à pontuação é que alguns títulos não teriam sido ganhados por aqueles que foram campeões. Isso não

tira o mérito de cada campeão mas não deixa de ser interessante observar que:

- Eddie Irvine teria sido campeão ao invés de Mika Hakkinen em 1999.
- Michael Schumacher teria sido campeão ao invés de Jacques Villenueve em 1997.
- Damon Hill teria sido campeão ao invés de Michael Schumacher em 1994.
- Alain Prost teria sido campeão ao invés de Ayrton Senna em 1988, ao invés de Niki Lauda em 1984 e ao invés de Nelson Piquet em 1983.
- Niki Lauda teria sido campeão ao invés de James Hunt em 1976.
- Graham Hill teria sido campeão ao invés de Jim Clark em 1965 e ao invés de John Surtees em 1964.

Além da padronização da pontuação, temos um outro problema: o da alta variância entre o número de corridas que os pilotos competiam no início da categoria e o número de corridas que competem agora. Assim, ao invés de utilizar todas as informações das carreiras dos pilotos para compará-los, dando um peso único para cada piloto independente do piloto a qual ele está sendo comparado, utilizamos somente as estatísticas de quando ambos os pilotos corriam juntos. Por exemplo, ao comparar Ayrton Senna e Michael Schumacher, ao invés de utilizar o número de vitórias totais que ambos possuem na carreira, é utilizado somente o número de corridas que eles ganharam enquanto ambos corriam juntos. Dessa forma, a discrepância do número de corridas entre as décadas não influenciarão no resultado final. As variáveis utilizadas na pesquisa são abordadas ao longo do texto.

Capítulo 3

Metodologia

Neste trabalho estamos interessados em identificar, a partir de uma regra de decisão pré-estabelecida, o melhor piloto de Fórmula 1 da história. Por se tratar de um problema em que há uma relação de interdependência entre os pilotos que disputaram os campeonatos de Fórmula 1, podemos utilizar um modelo baseado em grafos para representar e solucionar esse problema. Na primeira seção deste capítulo, apresentamos as principais noções relativas a teoria dos grafos, no qual utilizamos como referência base o livro de Newman (2018). Na seção seguinte, definimos cadeias de Markov e exploramos o conceito de distribuição estacionária. Para isso, utilizamos como referência o livro de Brémaud (2013) e o livro de Levin e Peres (2017). Por fim, apresentamos uma seção sobre construção e simulação de cadeias de Markov, na qual utilizamos como referência Ferrari e Galves (1997) e Häggström *et al.* (2002).

3.1 Grafos

Nessa forma de modelagem, os pilotos são denominado **nós** ou **vértices** e as relações de interdependência entre os pilotos são denominadas **arestas** ou **arcos**. No modelo de grafo que adotamos neste trabalho, denotamos por V o conjunto de todos os pilotos de Fórmula 1 e dizemos que dois pilotos $i \in V$ e $j \in V$ possuem uma relação de interdependência caso ambos já tenham corrido alguma corrida juntos. Assim, denotamos por E o subconjunto de $V \times V$ composto pelos pares de pilotos que possuem uma relação de interdependência. Em outras palavras, V é o conjunto de todos os vértices do nosso grafo e E é o conjunto de todas as arestas. Portanto, o par (V, E) define o grafo G que consideramos nos estudos.

Definição 3.1 *Um **grafo** G é um par (V, E) em que V é um conjunto de vértices (objetos em estudo) e E é um conjunto de arestas (conexões entre os objetos de estudo), isto é, $E \subset V \times V$.*

Quando as propriedades das relações entre os vértices do conjunto V não dependem de sua origem, os pares de vértices pertencentes ao conjunto E são não-ordenados, isto é, $(i, j) = (j, i)$ para todo $i, j \in V$. Nesse caso, o grafo $G = (V, E)$ é denominado **não-ordenado** e podemos representá-lo em um plano desenhando os vértices como pontos ou círculos e as arestas como traços que ligam dois pontos.

Exemplo 3.2 *Seja $V = \{a, b, c, d, e\}$ um conjunto com apenas 5 pilotos. Considere que os pilotos a e b correram na temporada 1, os pilotos b, c e d correram na temporada 2 e os pilotos d e e correram na temporada 3. Dessa forma, se dois pilotos possuem uma relação de interdependência quando correram juntos em uma mesmo temporada, então o conjunto E das arestas é dado por*

$$E = \{(a, b), (b, c), (b, d), (c, d), (d, e)\}. \tag{3.3}$$

Nesse caso, podemos utilizar o grafo $G = (V, E)$ para modelar a conexão entre os 5 pilotos em estudo. O grafo G pode ser representado de forma pictórica como a seguir.

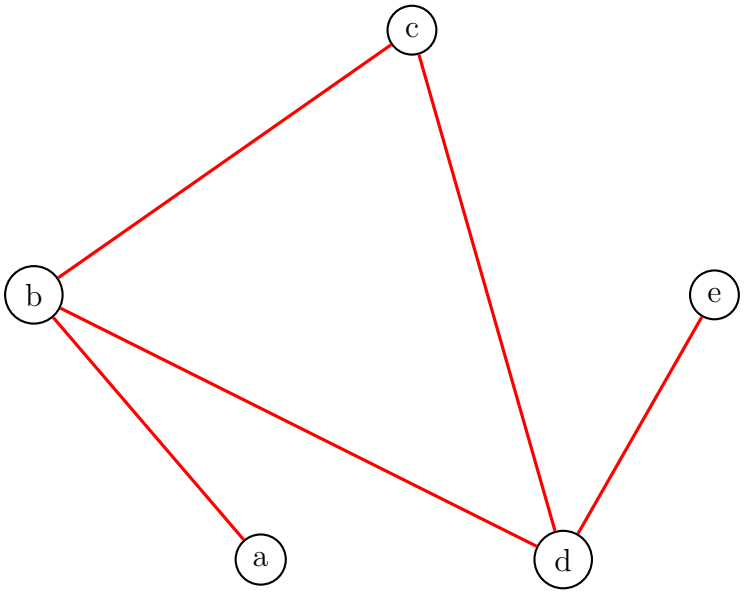

Figura 3.1: Representação pictórica do grafo $G = (V, E)$, em que $V = \{a, b, c, d, e\}$ e $E = \{(a, b), (b, c), (b, d), (c, d), (d, e)\}$.

Quando as relações de interdependência entre os vértices do conjunto V dependem da

sua origem, os pares de vértices pertencentes ao conjunto E são ordenados, isto é, $(i, j) \neq (j, i)$ para todo $i, j \in V$. Nesse caso, o grafo $G = (V, E)$ é denominado **direcionado** ou **orientado** e podemos representá-lo em um plano desenhando os vértices como pontos ou círculos e as arestas como setas que ligam dois pontos na direção da relação considerada.

Definição 3.4 *Um grafo $G = (V, E)$ é dito **orientado** quando o conjunto E de arestas define uma relação de ordem sobre o conjunto V de vértices.*

Um grafo pode conter informações associadas tanto aos seus vértices como às suas arestas. Essas informações podem ser tanto numéricas como alfabéticas e podem ser descritas no plano junto aos elementos no qual estão associadas. Tais informações são denominadas **rótulos** e o grafo que possui rótulos em suas arestas ou vértices é denominado **grafo rotulado**. No Exemplo 3.2, o grafo G é rotulado nos vértices, pois informações alfabéticas são atribuídas a eles.

Nesse sentido, grafos podem possuir ponderações ou pesos. No modelo de grafo que vamos utilizar para modelar a relação de interdependência entre os pilotos de Fórmula 1, estamos interessados em definir para cada par de pilotos um valor que represente o peso que um dos pilotos possui sobre a performance do outro, o que permite a criação de uma medida de comparação de suas performances. Medida essa que será essencial na especificação da regra de decisão a ser utilizada para o ranqueamento dos pilotos de Fórmula 1. Portanto, é razoável pensar que para cada par de pilotos, um peso diferente é atribuído à respectiva aresta.

Definição 3.5 *Quando a relação ordinária em um grafo é impactada por características do problema de forma a fortalecer ou enfraquecer a conexão entre os vértices, pode-se atribuir **pesos** entre os vértices. Nesse caso, o grafo $G = (V, E)$ é denominado **grafo com pesos** ou **grafo ponderado**.*

No problema que pretendemos modelar, a relação de interdependência entre dois pilotos $i \in V$ e $j \in V$ depende da origem da relação, ou seja, o grafo $G = (V, E)$ é orientado. Nesse sentido, o peso atribuído à aresta (i, j) pode ser diferente daquele atribuído à aresta (j, i) pois $(i, j) \neq (j, i)$. Além disso, se o piloto i é considerado melhor do que o piloto j, então o peso atribuído à aresta (i, j) será grande, consequentemente, o piloto j será considerado pior do que o piloto i e, então, o peso a ser atribuído à aresta (j, i) deverá ser pequeno. Como o peso da aresta (i, j) pode ser diferente do peso da aresta (j, i) para

todo $i, j \in V$, é razoável utilizar um grafo orientado com pesos para modelagem do nosso problema.

Exemplo 3.6 *Utilizando o mesmo conjunto V de 5 pilotos do Exemplo 3.2, vamos definir pesos para cada relação de interdependência pertencentes ao conjunto E a fim de entendermos como é a representação pictórica de um grafo orientado e com pesos. Para a definição dos pesos, podemos levar em consideração características dos pilotos em estudo. Para fins didáticos, vamos supor que os pesos a serem atribuídos a cada aresta seja uma função de uma dada variável. Digamos que a variável escolhida seja o número de vitórias dos pilotos em sua carreira. Suponha que o*

- *piloto a obteve* 10 *vitórias;*
- *piloto b obteve* 20 *vitórias;*
- *piloto c obteve* 30 *vitórias;*
- *piloto d obteve* 40 *vitórias;*
- *piloto e obteve* 50 *vitórias.*

Uma maneira de definir os pesos $\omega(i \to j)$ para cada par de pilotos $(i, j) \in E$ é

$$\omega(i \to j) = \frac{n^o \text{ de vitórias do piloto } j}{n^o \text{ de vitórias do piloto } i}, \tag{3.7}$$

isto é, a razão entre o número de vitórias do piloto j e o número de vitórias do piloto i. O número de vitórias de um piloto é uma estatística simples, porém descreve bem sua carreira, para um exemplo, é suficiente.

Dessa forma, obtemos os seguintes pesos:

- $\omega(a \to b) = \frac{20}{10} = 2$;
- $\omega(b \to a) = \frac{10}{20} = 0,5$;
- $\omega(b \to c) = \frac{30}{20} = 1,5$;
- $\omega(c \to b) = \frac{20}{30} = 0,66$;
- $\omega(b \to d) = \frac{40}{20} = 2$;
- $\omega(d \to b) = \frac{20}{40} = 0,5$;

- $\omega(c \to d) = \frac{40}{30} = 1,33$;
- $\omega(d \to c) = \frac{30}{40} = 0,75$;
- $\omega(d \to e) = \frac{50}{40} = 1,25$;
- $\omega(e \to d) = \frac{40}{50} = 0,8$.

Podemos representar de forma pictória o grafo G como na Figura 3.2. Note que agora utilizamos setas para representar as arestas, indicando a orientação da relação e sinalizamos os pesos ao longo das setas, representando os rótulos das arestas.

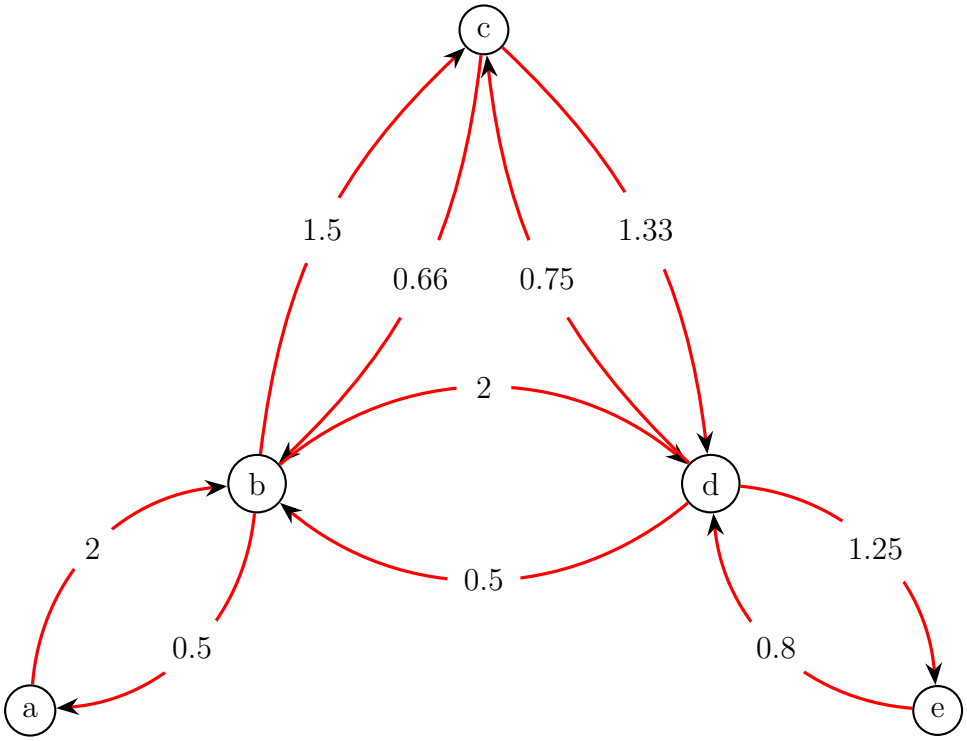

Figura 3.2: Representação pictória do grafo direcionado com pesos $G = (V, E)$, em que $V = \{a, b, c, d, e\}$ e $E = V \times V$. Os pesos foram definidos a partir de uma variável fictícia que representa o número de vitórias de cada piloto ao longo da história como definido no Exemplo 3.6.

A **ordem** de um grafo $G = (V, E)$ é a cardinalidade do seu conjunto V de vértices, o qual vamos denotar por $|V|$, enquanto o **tamanho** de um grafo G é a cardinalidade do seu conjunto E de arestas, que denotamos por $|E|$. Nesse sentido, um grafo é dito **finito** se possuir ordem e tamanho finitos. Caso contrário, dizemos que o grafo é **infinito**. Neste trabalho, o grafo G que vamos trabalhar no processo de modelagem do nosso problema tem como conjunto V de vértices o conjunto de todos os pilotos considerados pela base de dados, a qual encontra-se disponível no Kaggle.

A todo grafo é possível associar uma matriz A em que cada entrada descreve quais são os vértices que possuem relação de interdependência e qual é a força dessa relação. Denominamos essa matriz A de **matriz de adjacência**.

Exemplo 3.8 *A matriz de adjacência para o grafo não-direcionado do Exemplo 3.6 é a seguinte*

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}.$$

Note que essa matriz possui somente entradas binárias, i.e., ou iguais a 0 *ou iguais a* 1*, em que* 1 *simboliza a existência de uma relação de interdependência entre dois pilotos e* 0 *a não-existência.*

No caso do grafo G do nosso problema, a matriz de adjacência A é uma matriz de ordem $|V| \times |V|$ tal que o elemento a_{ij} que está na linha i e coluna j da matriz A é tal que

$$a_{ij} = \begin{cases} \omega(i \to j), & \text{se os pilotos } i \text{ e } j \text{ participaram de alguma corrida juntos,} \\ 0, & \text{caso contrário.} \end{cases}$$

em que $\omega(i \to j) \in \mathbb{R}$ representa o peso que o piloto i possui sobre a performance do piloto j, para todo $i, j = 1, 2, \ldots, |V|$.

Exemplo 3.9 *A matriz de adjacência para o grafo orientado com pesos do Exemplo 3.6 é a seguinte*

$$B = \begin{bmatrix} 0 & 2 & 0 & 0 & 0 \\ 0,5 & 0 & 1,5 & 2 & 0 \\ 0 & 0,66 & 0 & 1,33 & 0 \\ 0 & 0,5 & 0,75 & 0 & 1,25 \\ 0 & 0 & 0 & 0,8 & 0 \end{bmatrix}.$$

Nesse caso, a matriz não é mais binária e a existência de uma relação de interdependência entre os pilotos é estabelecida a partir do peso dessa conexão. Quanto maior esse peso mais forte é a relação de dependência entre os pilotos e quanto menor o peso, mais fraca

é essa relação.

Note que em ambos os casos, assumimos que $\omega(i \to i) = 0$ *para todo* $i \in V$ *e, portanto, os elementos da diagonal principal dessa matriz de adjacência são todos iguais a zero. Assumimos o valor 0 e não 1 pois não faz sentido existir comparação de um piloto com ele mesmo. Além disso, é possível observar que quando o grafo é não-direcionado a matriz de adjacência é simétrica e quando ele é não-direcionado ela não é simétrica.*

3.2 Cadeias de Markov

Processos estocásticos são modelos matemáticos que podem ser utilizados para descrever a evolução de uma ou mais variáveis aleatórias ao longo do tempo. Estes são utilizados em várias áreas do conhecimento para modelar estocasticamente fenômenos com dinâmica temporal. Dentro da área de processos estocásticos, as cadeias de Markov são uma classe importante e amplamente utilizada. Estas descrevem a evolução de um sistema ao longo do tempo, de forma que a probabilidade de transição para qualquer estado futuro depende apenas de seu estado presente e, possivelmente, de um número finito de estados do passado. Nesta monografia, vamos explorar a aplicação das cadeias de Markov em um problema específico: ranquear pilotos da Fórmula 1. Como mencionamos na seção anterior, cada piloto considerado no estudo é representado por um vértice de um grafo direcionado com pesos, de forma que dois vértices quaisquer desse grafo são interligados por uma aresta, caso os respectivos pilotos tenham corrido juntos em algum momento ao longo da história. Os pesos dessas arestas são definidos com base em um conjunto de variáveis, as quais discutimos em mais detalhes no próximo capítulo. O ranqueamento dos pilotos é feito através da simulação de cadeias de Markov, cujas probabilidades de transição são definidas a partir da matriz de adjacência do grafo em estudo. O objetivo é ranquear os pilotos de acordo com o número de visitas da cadeia, sendo que quanto mais visitas o vértice recebeu, melhor classificado é o piloto associado.

Definição 3.10 *Um* ***processo estocástico*** *a tempo discreto é uma sequência de variáveis aleatórias* $\boldsymbol{X} := \{X_t : t \in \mathbb{T}\}$ *definidas em algum espaço de probabilidade* $(\Omega, \mathcal{F}, \mathbb{P})$*, em que* $\mathbb{T}$ *é um conjunto de índices enumerável.*

Os possíveis valores que as variáveis aleatórias X_t, $t \in \mathbb{T}$, podem assumir são chamados de **estados**. O conjunto de todos os estados possíveis é chamado de **espaço**

de estados. Em nosso problema, vamos assumir, sem perda de generalidade, que $\mathbb{T} = \mathbb{N} := \{0, 1, 2, \ldots\}$. Nesse caso, dizemos que o processo estocástico $\boldsymbol{X}$ é um **processo estocástico a tempo discreto** e preferimos escrever X_n ao invés de X_t. Além disso, assumimos que o espaço de estados é o conjunto V dos vértices do grafo G que utilizamos no processo de modelagem, ou seja, o conjunto de todos os pilotos do banco de dados que são considerados no estudo.

Definição 3.11 *Um processo estocástico a tempo discreto* $\boldsymbol{X} := \{X_n : n \in \mathbb{N}\}$ *definido em um espaço de probabilidade* $(\Omega, \mathcal{F}, \mathbb{P})$ *assumindo valores no espaço de estado* V *é dito ser* ***homogêneo*** *quando suas probabilidades de transição são invariantes ao longo do tempo, ou seja, para todo número natural* n*,*

$$\mathbb{P}\left(X_n = b \middle| \bigcap_{i=1}^{n} \{X_{n-i} = a_{n-i}\}\right) = \mathbb{P}\left(X_{n+1} = b \middle| \bigcap_{i=1}^{n} \{X_{n+1-i} = a_{n-i}\}\right), \tag{3.12}$$

quaisquer que sejam $a_0, \ldots, a_{n-1}, b \in V$*.*

A suposição de homogeneidade é razoável para nosso modelo, pois a regra de decisão que utilizamos para definir a dinâmica da cadeia depende apenas do par de pilotos comparados e não do instante de tempo no qual os pilotos são visitados. Essa é uma simplificação importante e torna a definição das probabilidades de transição muito mais fácil.

Definição 3.13 *Seja* $\boldsymbol{X} := \{X_n : n \in \mathbb{N}\}$ *um processo estocástico a tempo discreto definido em um espaço de probabilidade* $(\Omega, \mathcal{F}, \mathbb{P})$ *e assumindo valores no espaço de estados* V*. Dizemos que* $\boldsymbol{X}$ *possui a* ***propriedade de Markov*** *quando para todo número natural* n*,*

$$\mathbb{P}\left(X_{n+1} = b | X_n = a, X_{n-1} = a_{n-1}, \ldots, X_0 = a_0\right) = \mathbb{P}\left(X_{n+1} = b | X_n = a\right), \tag{3.14}$$

quaisquer que sejam $a_0, a_1, \ldots, a_{n-1}, a, b \in V$*.*

A propriedade de Markov afirma que dado o estado atual do processo estocástico, a probabilidade de sua transição para outro estado futuro é independente do histórico passado do processo. Uma cadeia que possui essa propriedade é denominada **cadeia de Markov de ordem** 1. Neste trabalho, vamos nos referir a elas simplesmente como **cadeias de Markov**.

As transições de uma cadeia de Markov, assumindo valores em um espaço de estados

discreto V finito cuja cardinalidade é $|V|$, podem ser representadas matricialmente por uma matriz de ordem $|V| \times |V|$ cujas entradas são as probabilidades de transição. Essa matriz é denominada matriz de transição.

Definição 3.15 *Seja* $\boldsymbol{X} := \{X_n : n \in \mathbb{N}\}$ *uma cadeia de Markov a tempo discreto definida em um espaço de probabilidade* $(\Omega, \mathcal{F}, \mathbb{P})$ *e assumindo valores no espaço de estados* V*. Denominamos* ***matriz de transição*** *da cadeia* $\boldsymbol{X}$ *à matriz* P *de ordem* $|V| \times |V|$ *tal que para todo número natural* n*,*

$$P\left(a, b\right) := \mathbb{P}\left(X_{n+1} = b | X_n = a\right), \tag{3.16}$$

quaisquer que sejam $a, b \in V$*, em que* $P(a, b) \in [0, 1]$ *é o elemento da matriz* P *que está na linha correspondente ao estado* a *e na coluna correspondente ao estado* b*. Para que* P *seja uma matriz de transição, é necessária a seguinte propriedade:*

$$\sum_{b \in V} P(a, b) = 1. \tag{3.17}$$

Assumimos a Markovianidade para nosso problema pois, em sua construção, as comparações de cada piloto são feitas duas a duas. Portanto, a cadeia não leva em consideração o histórico de pilotos visitados pela cadeia.

A matriz de transição P pode ser representada por seu **grafo de transição**, de forma que os vértices são os estados de V. Tal grafo possui um elo orientado de a para b se, e somente se, $P(a, b) > 0$. Neste caso, o peso de cada elo é dado por $P(a, b)$.

Exemplo 3.18 *Considere o mesmo conjunto* V *de 5 pilotos do Exemplo 3.2, no qual definimos os pesos* ω *para a relação de interdependência entre os pilotos. Podemos definir as probabilidades de transição a partir de tais pesos da seguinte maneira:*

$$P(a, b) := \frac{\omega(a \to b)}{\sum_{b \in V} \omega(a \to b)}.$$

Dessa forma, a matriz de transição P é dada por

$$P = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0,125 & 0 & 0,375 & 0,5 & 0 \\ 0 & 0,33 & 0 & 0,67 & 0 \\ 0 & 0,2 & 0,3 & 0 & 0,5 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}.$$

A matriz de transição P pode ser vista como uma matriz de adjacência normalizada. Assim, o grafo de transição possui a mesma representação pictórica que o grafo do Exemplo 3.6, porém agora os pesos são dados pelas probabilidades de transição especificadas na matriz de transição P, conforme mostra a figura 3.3.

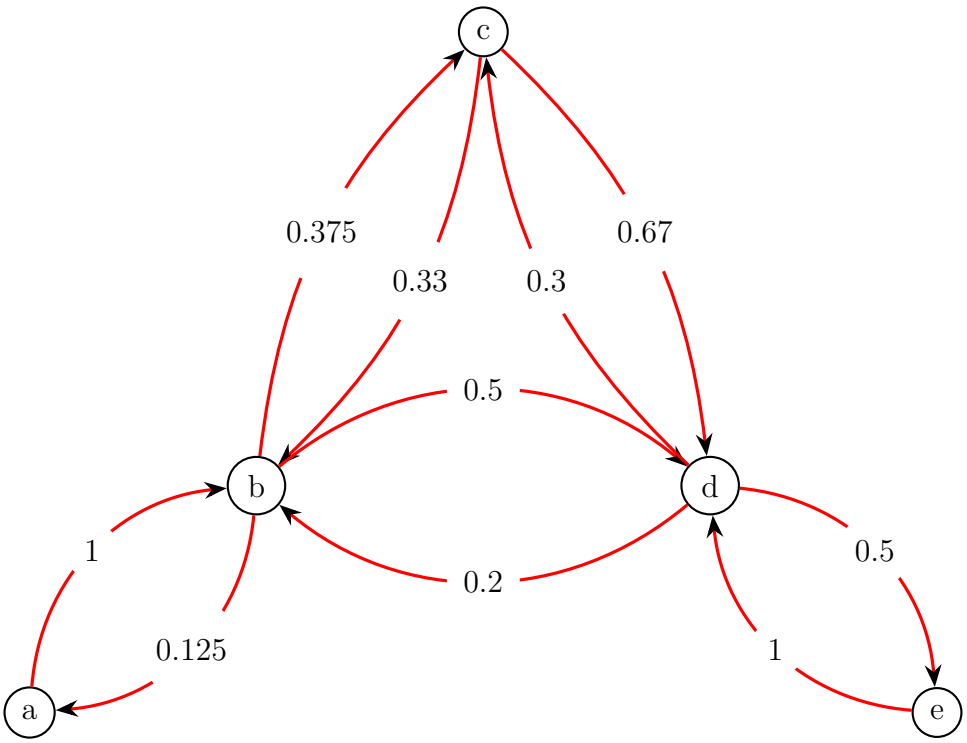

Figura 3.3: Representação pictórica do grafo de transição $G = (V, E)$, em que $V = \{a, b, c, d, e\}$ e $E = V \times V$ e as probabilidades de transição são dadas pela matriz P.

Definição 3.19 *Seja $\boldsymbol{X} := \{X_n : n \in \mathbb{N}\}$ uma cadeia de Markov a tempo discreto definida em um espaço de probabilidade $(\Omega, \mathcal{F}, \mathbb{P})$ e assumindo valores no espaço de estados V. O **estado inicial** da cadeia $\boldsymbol{X}$ é o valor assumido pela variável aleatória X_0, cuja distribuição é dada por um vetor $\boldsymbol{\mu}_0$ de ordem $|V|$ tal que cada entrada é dada por*

$$\mu_0(a) := \mathbb{P}(X_0 = a),\ a \in V. \tag{3.20}$$

Essa distribuição é denominada ***distribuição inicial****.*

No instante de tempo $n \in \mathbb{N}$, a distribuição da cadeia $\boldsymbol{X}$ pode ser descrita por um vetor $\boldsymbol{\mu}_n$ de ordem $|V|$ tal que

$$\mu_n(a) := \mathbb{P}(X_n = a),\ a \in V. \tag{3.21}$$

O vetor $\boldsymbol{\mu}_n$ é denominado **vetor de distribuição no tempo** n.

O interessante ao trabalhar com cadeias de Markov homogêneas é que podemos facilmente encontrar o vetor $\boldsymbol{\mu}_n$ de distribuição no tempo n da cadeia de Markov $\boldsymbol{X}$, a partir de um vetor de distribuição inicial $\boldsymbol{\mu}_0$ e de uma matriz de transição P. De fato, para todo $b \in V$

$$\begin{aligned}
\mu_n(b) &= \mathbb{P}\left(X_n = b\right) \\
&= \sum_{a \in V} \mathbb{P}\left(X_{n-1} = a, X_n = b\right) \\
&= \sum_{a \in V} \mathbb{P}\left(X_n = b | X_{n-1} = a\right) \mathbb{P}\left(X_{n-1} = a\right) \\
&= \sum_{a \in V} P(a,b)\mu_{n-1}(a).
\end{aligned}$$

Em forma matricial, $\boldsymbol{\mu}_n = P\boldsymbol{\mu}_{n-1}$. Logo, por recursividade, obtemos $\boldsymbol{\mu}_n = P^n\boldsymbol{\mu}_0$.

A igualdade $\boldsymbol{\mu}_n = P\boldsymbol{\mu}_{n-1}$ também evidencia que para encontrar a distribuição no tempo $n+m$ a partir do tempo n, basta multiplicar o vetor de distribuição $\boldsymbol{\mu}_n$ pela matriz de transição m vezes, i.e., P^m. Assim, podemos definir a **matriz de transição associada a** m **passos no futuro** de modo que a entrada associada a transição do piloto $a \in V$ para o piloto $b \in V$ por

$$P^m(a,b) = \mathbb{P}(X_{n+m} = b | X_n = a).$$

Exemplo 3.22 *Continuando com o Exemplo 3.2 dos 5 pilotos, é possível, agora, calcular as matrizes de transição para os primeiros 5 passos. Lembramos que a matriz de transição* P *é*

$$P = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0,125 & 0 & 0,375 & 0,5 & 0 \\ 0 & 0,33 & 0 & 0,67 & 0 \\ 0 & 0,2 & 0,3 & 0 & 0,5 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}.$$

Para 2 e 5 passos no futuro temos, respectivamente

$$P^2 = \begin{bmatrix} 0,12 & 0 & 0,38 & 0,5 & 0 \\ 0 & 0,35 & 0,15 & 0,25 & 0,25 \\ 0,04 & 0,13 & 0,32 & 0,16 & 0,34 \\ 0,03 & 0,1 & 0,08 & 0,80 & 0 \\ 0 & 0,2 & 0,3 & 0 & 0,5 \end{bmatrix}.$$

$$P^5 = \begin{bmatrix} 0,01 & 0,22 & 0,19 & 0,31 & 0,26 \\ 0,03 & 0,14 & 0,18 & 0,50 & 0,16 \\ 0,02 & 0,15 & 0,14 & 0,57 & 0,12 \\ 0,02 & 0,20 & 0,25 & 0,18 & 0,35 \\ 0,03 & 0,12 & 0,11 & 0,69 & 0,05 \end{bmatrix}.$$

Para efeitos de interpretação, o valor de maior probabilidade da matriz de transição em 5 passos $P(d, e) = 0.69$ é a probabilidade de transição do piloto d para o piloto e após 5 passos, ou seja, se no tempo n a cadeia estava no piloto d, temos 69% de chance de que, no tempo $n + 5$, a cadeia se encontre em e.

Definição 3.23 *Um estado $b \in V$ é dito ser **acessível** por um estado $a \in V$ se existe $M \geq 0$ tal que $P^M(a, b) > 0$, ou seja, em um número finito de passos, é possível partir de a e chegar em b. A notação utilizada para expressar acessibilidade é $a \to b$. Falamos também que a e b se **comunicam** se a é acessível a partir de b e b é acessível a partir de a. A notação utilizada para expressar comunicabilidade é $a \leftrightarrow b$.*

A relação de comunicação gera partições disjuntas de classes de equivalência do espaço de estado V chamadas de **classes de comunicação**.

Definição 3.24 *Seja F o conjunto de estados tal que para todo $a \in F$,*

$$\sum_{b \in F} P(a,b) = 1.$$

Os estados pertencentes a F são chamados de ***estados fechados****. Se em uma cadeia de Markov existe somente uma classe de comunicação, então, tanto a cadeia quanto a respectiva matriz de transição e o grafo de transição são denominados* ***irredutíveis****.*

Por mais que estejamos interessados na comparação entre pares de pilotos, o objetivo final é ranquear todos os pilotos de Fórmula 1, a partir do número de visitas de cadeias de Markov. Para que todos os pilotos sejam ranqueados, é necessário que qualquer piloto seja acessível a partir de outro, ou seja, que todos os pilotos se comuniquem. Portanto, é importante que as cadeias a serem utilizadas em nosso trabalho sejam irredutíveis.

Exemplo 3.25 *Para o Exemplo 3.2 dos 5 pilotos, a cadeia é irredutível. Neste caso, partindo de qualquer piloto, existe uma probabilidade positiva da cadeia acessar qualquer outro piloto após 5 passos. De fato,*

- *Piloto A:*

$$(1,0,0,0,0)\begin{bmatrix} 0,01 & 0,22 & 0,19 & 0,31 & 0,26 \\ 0,03 & 0,14 & 0,18 & 0,50 & 0,16 \\ 0,02 & 0,15 & 0,14 & 0,57 & 0,12 \\ 0,02 & 0,20 & 0,25 & 0,18 & 0,35 \\ 0,03 & 0,12 & 0,11 & 0,69 & 0,05 \end{bmatrix} = (0,01, 0,22, 0,19, 0,31, 0,26)\,;$$

- *Piloto B:*

$$(0,1,0,0,0)\begin{bmatrix} 0,01 & 0,22 & 0,19 & 0,31 & 0,26 \\ 0,03 & 0,14 & 0,18 & 0,50 & 0,16 \\ 0,02 & 0,15 & 0,14 & 0,57 & 0,12 \\ 0,02 & 0,20 & 0,25 & 0,18 & 0,35 \\ 0,03 & 0,12 & 0,11 & 0,69 & 0,05 \end{bmatrix} = (0,03, 0,14, 0,18, 0,5, 0,16)\,;$$

- *Piloto C:*

$$(0,0,1,0,0)\begin{bmatrix} 0,01 & 0,22 & 0,19 & 0,31 & 0,26 \\ 0,03 & 0,14 & 0,18 & 0,50 & 0,16 \\ 0,02 & 0,15 & 0,14 & 0,57 & 0,12 \\ 0,02 & 0,20 & 0,25 & 0,18 & 0,35 \\ 0,03 & 0,12 & 0,11 & 0,69 & 0,05 \end{bmatrix} = (0,02,0,15,0,14,0,57,0,12)\,;$$

- *Piloto D:*

$$(0,0,0,1,0)\begin{bmatrix} 0,01 & 0,22 & 0,19 & 0,31 & 0,26 \\ 0,03 & 0,14 & 0,18 & 0,50 & 0,16 \\ 0,02 & 0,15 & 0,14 & 0,57 & 0,12 \\ 0,02 & 0,20 & 0,25 & 0,18 & 0,35 \\ 0,03 & 0,12 & 0,11 & 0,69 & 0,05 \end{bmatrix} = (0,02,0,2,0,25,0,18,0,35)\,;$$

- *Piloto E:*

$$(0,0,0,0,1)\begin{bmatrix} 0,01 & 0,22 & 0,19 & 0,31 & 0,26 \\ 0,03 & 0,14 & 0,18 & 0,50 & 0,16 \\ 0,02 & 0,15 & 0,14 & 0,57 & 0,12 \\ 0,02 & 0,20 & 0,25 & 0,18 & 0,35 \\ 0,03 & 0,12 & 0,11 & 0,69 & 0,05 \end{bmatrix} = (0,03,0,12,0,11,0,69,0,05)\,;$$

Podemos perceber que, após multiplicar a distribuição inicial, degenerada no estado inicial, por P^5 obtemos um vetor cujos elementos são todos maiores do que zero. Concluímos, portanto, que em 5 passos qualquer piloto é acessível a partir de qualquer outro e, consequentemente, a cadeia possui apenas uma classe de comunicação sendo, dessa forma, irredutível.

Teorema 3.26 *Para qualquer cadeia de Markov irredutível com matriz de transição P é possível encontrar uma única partição $C_0, C_1, \ldots, C_{d-1}$ do espaço de estados V tal que*

para todo k *e todo* $a \in C_k$,

$$\sum_{b \in C_{k+1}} P(a,b) = 1.$$

Por convenção, $C_0 = C_d$ *e* d *é maximal, ou seja, não existe outra partição entre* d' *classes tal que* $d' > d$.

Em outras palavras, a cadeia transiciona entre cada classe, a cada passo, de C_k para C_{k+1}, para todo $k \in \{0, 1, ..., d-1\}$.

Demonstração. *A prova do Teorema 3.26 pode ser encontrado em Levin e Peres (2017).*

□

Se pensarmos cada C_i como o conjunto de pilotos que correram na temporada de número i e reparar que em nenhum ano todos os pilotos foram substituídos ao fim da temporada, conseguimos criar uma separação da população tal que as preposições do teorema 3.26 são válidas. Dessa forma, podemos considerar o grafo que trabalhamos irredutível.

Definição 3.27 *O número* $d \geq 1$ *de classes mencionado no Teorema 3.26 é chamado de* ***período*** *da cadeia. As* d *classes* $C_0, C_1, \ldots, C_{d-1}$ *são chamadas de* ***classes cíclicas****.*

Outra forma de definir o conceito de período é pelo tempo que demora para a cadeia sair de um estado e retornar até ele mesmo. Seja

$$\tau(a) := \{n \geq 1 : P^n(a,a) > 0\} \tag{3.28}$$

o conjunto de tempos no qual é possível uma cadeia de Markov retornar para sua posição inicial a. O período de um estado a é definido como o maior divisor comum de $\tau(a)$.

Definição 3.29 *Quando um estado possui período 1, o denominamos* ***aperiódico****, caso contrário, o denominamos* ***periódico****.*

Todos os estados de uma cadeia são aperiódicos, dizemos que a cadeia é aperiódica. A suposição de aperiodicidade é razoável para o problema de ranqueamento dos pilotos pois, para a construção do *ranking*, assumimos que existe uma única distribuição estacionária, i.e., uma única distribuição que seja invariante por translação temporal.

Em linha com a separação comentada anteriormente, ao pensarmos que todos os C_i se auto-acessam (possuem período 1), podemos considerar o grafo que trabalhamos aperiódico.

Exemplo 3.30 *Considere o grafo de transição do Exemplo 3.2.*

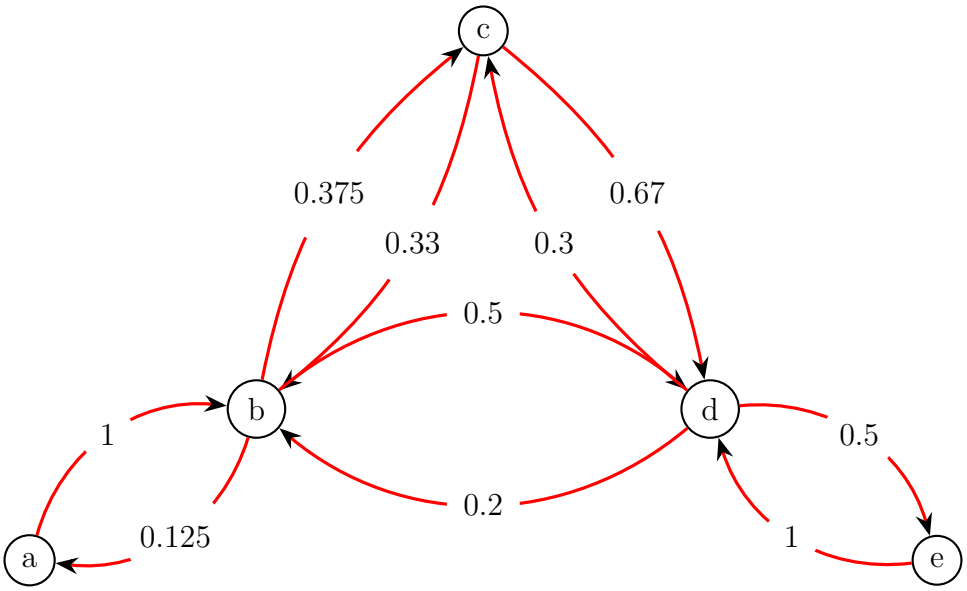

Figura 3.4: Representação pictórica do grafo de transição $G = (V, E)$, em que $V = \{a, b, c, d, e\}$ e $E = V \times V$.

Vimos anteriormente que tal cadeia é irredutível, então basta encontrarmos um estado aperiódico para mostrar que a cadeia do exemplo também é aperiódica. Peguemos o estado b como exemplo. Sabemos que dois caminhos possíveis de retorno são

- $b \to d \to b$;
- $b \to c \to d \to b$.

Assim, temos

$$\tau(b) = \{2, 3, ...\}$$

e, portanto, o maior divisor comum de $\tau(b)$ é 1. Logo, o estado b e, por consequência, a cadeia do exemplo é aperiódica.

Definição 3.31 *A distribuição de probabilidade π que satisfaz*

$$\pi = \pi P \tag{3.32}$$

é chamada de ***distribuição estacionária*** *da matriz de transição P, ou da cadeia de Markov correspondente.*

A Equação 3.32 é chamada de **equação total do balanço** e nos diz que se uma cadeia iniciou com a distribuição estacionária ($\mu_0 = \pi$) ou alcançou em algum momento a distribuição estacionária ($\mu_n = \pi$), então ela não dependerá de n e manterá a mesma distribuição para sempre. Neste caso, chamamos a cadeia de **estacionária**.

A distribuição estacionária é de suma importância para o trabalho pois, ao encontrar a medida estacionária, a utilizamos como método de decisão para o ranqueamento dos pilotos.

Teorema 3.33 *Dada uma cadeia de Markov $\boldsymbol{X} := \{X_n : n \in \mathbb{N}\}$ irredutível e aperiódica a tempo discreto definida em um espaço de probabilidade $(\Omega, \mathcal{F}, \mathbb{P})$ e assumindo valores no espaço de estados V com matriz de transição P, então, existe uma única distribuição de probabilidade π tal que*

$$\pi P = \pi.$$

Demonstração. *Prova do teorema no apêndice A.*

□

Como mencionamos, vamos utilizar a distribuição estacionária para ranquear os pilotos. Nesse sentido, além de garantir a existência e a unicidade de tal distribuição, precisamos garantir que o passeio da cadeia de Markov sobre o grafo de pilotos convergirá para a cadeia estacionária, após uma quantidade finita de passos.

Definição 3.34 *A* ***distância em variação total (DVT)*** *entre duas distribuições de probabilidade μ e ν em V é definida como o máximo das diferenças entre μ e ν quando aplicado em um evento $A \subset V$, ou seja,*

$$||\mu - \nu||_{VT} = \max_{A \subset V} |\mu(A) - \nu(A)|. \tag{3.35}$$

Teorema 3.36 *Suponha P irredutível e aperiódica com distribuição estacionária π. Então existe constantes $\alpha \in (0, 1)$ e $C > 0$ tal que*

$$\max_{a \in V} ||P^n(a, .) - \pi||_{VT} \leq C\alpha^n. \tag{3.37}$$

Demonstração. *Prova do teorema no apêndice B.*

□

No Teorema 3.36, quando $n \to \infty$, a constante α elevada a n converge para 0 e, com isso, P^n converge para Π sendo Π uma matriz estocástica cujas linhas são iguais ao vetor $\boldsymbol{\pi}$. Isso comprova que, em nosso modelo, para um n suficientemente grande, chegaremos em uma única distribuição estacionária, a qual é utilizada para ranquear os pilotos do nosso conjunto de dados.

Nas simulações descritas posteriormente, utilizamos $C\alpha^n = 0,000001$.

Exemplo 3.38 *Voltando ao exemplo dos 5 pilotos, podemos facilmente encontrar a distribuição estacionária da cadeia. Ao calcular o valor da matriz de transição em um tempo muito distante, como 100 passos, por exemplo, temos um resultado bem interessante:*

$$P^{100} = \begin{bmatrix} 0,0208 & 0,1660 & 0,1874 & 0,4172 & 0,2086 \\ 0,0208 & 0,1660 & 0,1874 & 0,4172 & 0,2086 \\ 0,0208 & 0,1660 & 0,1874 & 0,4172 & 0,2086 \\ 0,0208 & 0,1660 & 0,1874 & 0,4172 & 0,2086 \\ 0,0208 & 0,1660 & 0,1874 & 0,4172 & 0,2086 \end{bmatrix}.$$

Para cadeias irredutíveis e aperiódicas, para n suficientemente grande P^n converge para uma matriz em que cada linha é a distribuição estacionária da cadeia. Para comprovar isso, basta aplicarmos os valores encontrados na equação total do balanço. Seja

$$\pi = (0,0208; 0,1660; 0,1874; 0,4172; 0,2086).$$

Ao multiplicar π por P, temos

$$\pi \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0,125 & 0 & 0,375 & 0,5 & 0 \\ 0 & 0,33 & 0 & 0,67 & 0 \\ 0 & 0,2 & 0,3 & 0 & 0,5 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix} = (0,0208; 0,1660; 0,1874; 0,4172; 0,2086) = \pi.$$

Neste caso, por estarmos trabalhando com uma matriz de transição pequena, conseguimos encontrar a distribuição estacionária para a cadeia e, assim, estabelecer o *ranking*.

Em ordem do melhor para o pior, temos os pilotos d, e, c, b, a. É interessante notar que, contra-intuitivamente, o piloto d está classificado como melhor que o piloto e, isso se deve ao fato de que o piloto d não possui um número de vitórias tão inferior ao e e, ao mesmo tempo, correu contra outros pilotos e possui um considerável maior número de vitórias em relação aos pilotos b e c. Além disso, o fato do piloto e ter corrido apenas contra o piloto a não faz muito sentido e pode ter influenciado o resultado do exemplo. Tal problema não está presente no conjunto de dados que analisamos neste trabalho.

Capítulo 4

Discussão do Problema

Seguindo a metodologia descrita no capítulo anterior, é necessário explicitar como construímos nosso modelo. Neste capítulo, discutimos como construímos a função f, o peso ω e a probabilidade de transição P que já foram mencionados anteriormente de forma que a cadeia tenda a visitar aquele que, de acordo com f, tenha a melhor performance perante seus pilotos contemporâneos. Após tal construção, são apresentados de forma breve alguns tópicos que são desenvolvidos no trabalho.

4.1 Enunciado do Problema

Cada piloto considerado na análise estatística representa um vértice de um grafo. Se dois pilotos já correram juntos em uma mesma temporada, então existe uma aresta conectando esses dois vértices. Assim, dois pilotos estão relacionados se, e somente se, tiverem sido oponentes em alguma temporada. A força dessa relação é definida a partir de uma medida de similaridade que, com base em um conjunto de covariáveis, informa o quão próxima é a performance dos pares de pilotos. Note que, dessa forma, estamos levando em consideração o efeito temporal das temporadas sobre a performance dos pilotos, uma vez que estamos mensurando a similaridade apenas entre pilotos que competiram juntos em uma ou mais temporadas. A regra que utilizamos para decidir qual é o melhor piloto da história é baseada em passeios aleatórios via uma matriz cujas probabilidades de transição entre dois vértices são funções das medidas de similaridade entre eles. Cada passeio aleatório iniciará sua trajetória em um vértice diferente do grafo e, após uma quantidade suficientemente grande de transições, analisamos quais foram os vértices mais visitados e, a partir deste número de visitas, identificamos quais são os melhores pilotos de todos os

tempos.

As variáveis que escolhemos, além de serem utilizadas em muitos *rankings* de forma subjetiva por páginas relevantes no mundo do automobilismo como a AutoSport (Jeffries, 2023), a TopGear (Barlow, 2022) e a Grande Prêmio (GP, 2022), descrevem os feitos de cada piloto e sua consistência na obtenção de tais números. As variáveis em questão são: número de títulos, número de vitórias, número de pódios e número de pontos padronizados.

O maior objetivo de qualquer piloto no início de uma temporada é ganhar o título ao final do ano. Logo, o **número de títulos** conquistados por um piloto mostra qual é sua efetividade em conquistar aquilo que todos almejam. Para a conquista de títulos, é necessário constantemente chegar nas primeiras posições das corridas. O **número de vitórias** conquistadas diz quantas vezes o piloto chegou na primeira posição, o **número de pódios** conquistados mostra quantas vezes o piloto chegou nas 3 primeiras posições, o **número de pontos padronizados** conquistados informa de forma ponderada qual foi a consistência do piloto entre os 10 primeiros lugares.

Seja $f : V^2 \to \mathbb{R}$ a função que quantificará a performance de um piloto $a \in V$ enquanto ele corria com outro piloto $b \in V$. Por exemplo, podemos assumir que

$$f(a, b) = \sum_{i=1}^{4} \beta_i x_i^{(a,b)},$$

i.e., f é uma função linear em que $x_i^{(a,b)}$ é tal que

- $x_1^{(a,b)}$: Número de títulos do piloto a enquanto corria com o piloto b,
- $x_2^{(a,b)}$: Número de vitórias do piloto a enquanto corria com o piloto b,
- $x_3^{(a,b)}$: Número de pódios do piloto a enquanto corria com o piloto b,
- $x_4^{(a,b)}$: Número de pontos padronizados do piloto a enquanto corria com o piloto b,
- $\beta_i \in \mathbb{R}$: Coeficientes que mensuram o peso de cada um dos respectivos $x_i^{(a,b))}$.

A função é o que quantifica o quão bom é o piloto e, por esse motivo, é necessário muita atenção em seu refinamento. Ao longo da discussão dos resultados, diferentes funções são testadas e seus respectivos resultados analisados para entender quais os efeitos que as alterações implicam no ranking final.

Todas as variáveis elencadas são quantitativas e afetam positivamente cada piloto, ou seja, quanto maior for o valor de $x_i^{(a,b)}$, $i \in \{1,2,3,4\}$, melhor o piloto avaliado será classificado. Ao definirmos os $\beta_i > 0$ para todo $i \in \{1,2,3,4\}$, a função $f(a,b)$ tem a característica de quantificar quão boa foi a carreira de a enquanto corria com b. O modo como serão definidos os coeficientes é discutido em mais detalhes no capítulo, porém, um modo simples e natural é explicitado no exemplo 4.1.

A função f é utilizada para definir os pesos no grafo de pilotos. Seja $\omega(a \to b)$ o peso que o piloto $a \in V$ possui sobre a performance do piloto $b \in V$ e defina

$$\omega(a \to b) := \frac{f(b,a)}{f(a,b)}.$$

Neste caso, $\omega(a \to b)$ é a razão entre a função de performance do piloto b no período que correu com o piloto a e da função de performance do piloto a no período que correu com o piloto b. O peso $\omega(a \to b)$ está bem definido, pois informa qual piloto entre os dois comparados é o melhor, e o quão melhor ele é:

- caso a seja melhor do que b, temos $f(a,b) > f(b,a)$ e, por consequência, $0 < \omega(a \to b) < 1$;
- caso a performance de a seja parecida com a de b, então $f(b,a)$ e $f(a,b)$ serão próximos e $\omega(a \to b)$ será próximo de 1;
- caso a seja pior do que b, temos $f(a,b) < f(b,a)$ e, por consequência, $\omega(a \to b) > 1$.

Para dois pilotos a e b, o peso $\omega(a \to b)$, quando dividido pela soma de todos os pesos daqueles pilotos que acessam b, define a probabilidade de transição da cadeia. Seja

$$P(a,b) = \frac{\omega(a \to b)}{\sum_{b \in V} \omega(a \to b)}$$

a probabilidade de transição do piloto $a \in V$ para o piloto $b \in V$. O peso $\omega(a \to b)$ definirá qual é a grandeza da probabilidade de transição de a até b. Portanto, quando $\omega(a \to b)$ é grande, isso significa que b possui uma superioridade sobre a, ou seja, b é melhor do que a. Ao mesmo tempo, a probabilidade de transição $P(a,b)$ não pode ser definida somente de acordo com a comparação entre a e b, é necessário levar em consideração a grandeza dos pesos de todos os outros pilotos que possuem relação com b. Por exemplo, dados três pilotos c, d, e, se o piloto c é pior do que o piloto d, porém muito pior do que o piloto e, é

necessário que, partindo de c, a cadeia tenda a ir para e com maior probabilidade que de ir para d. Dessa forma, a cadeia tenderá a visitar com maior frequência sempre o melhor piloto.

Vamos supor algumas situações num grafo entre três pilotos a, b e c para exemplificar a lógica por trás da estruturação da probabilidade de transição, fazendo com que a cadeia tenda a sempre visitar o piloto que, de acordo com a função f e o peso ω, é considerado o melhor, ou pelo menos melhor dentre aqueles que ele possui alguma conexão.

- O piloto a é melhor do que o piloto b e a é melhor do que o piloto c. Digamos que o piloto a é bem melhor do que b e a performance do piloto a é próxima à performance do piloto c.

  Neste caso, temos $\omega(a \to b) < 1$ (próximo de 0) e $\omega(a \to c) < 1$ (próximo de 1) e, portanto, a cadeia tende a ir para c quando está em a.

- O piloto a é melhor do que o piloto b e a é pior do que o piloto c.

  Neste caso, temos $\omega(a \to b) < 1$ e $\omega(a \to c) > 1$ e, portanto, a cadeia tende a ir para c quando está em a.

- O piloto a é pior do que o piloto b e a é melhor do que o piloto c.

  Neste caso, temos $\omega(a \to b) > 1$ e $\omega(a \to c) < 1$ e, portanto, a cadeia tende a ir para b quando está em a.

- O piloto a é pior do que o piloto b e a é pior do que o piloto c. Digamos que o piloto a é bem pior do que b e a performance do piloto a é próxima à performance do piloto c.

  Neste caso, temos $\omega(a \to b) > 1$ (muito maior do que 1) e $\omega(a \to c) > 1$ (próximo de 1) e, portanto, a cadeia tende a ir para b quando está em a.

A dificuldade em identificar o melhor piloto de Fórmula 1 da história está atrelada ao fato de que o desempenho observável de um piloto depende tanto do seu talento quanto de outras variáveis externas, tal como a competitividade de seus rivais. Além disso, existe uma dificuldade natural em comparar atletas de temporadas de épocas distintas, uma vez que as próprias regras, regulamento e maquinário utilizado sofreram e ainda sofrem mudanças ao longo do tempo. A forma que encontramos de superar essa dificuldade foi considerar anos comparáveis, para manter o maior número de regras e regulamentos iguais

ou o mais padronizado possível. Neste trabalho, usamos uma metodologia baseada em grafos que busca desconsiderar tais efeitos que os campeonatos possuem sobre a performance dos pilotos, diferente de outros ranqueamentos que sempre levam em consideração todas as estatísticas da carreira do piloto e os comparam todos juntos sem distinção das diferentes épocas. Desta forma, acreditamos que vamos obter uma comparação mais justa e sem perda da objetividade.

Exemplo 4.1 *Para exemplificar toda a descrição de como é construída a matriz de transição do modelo, utilizamos os 5 pilotos citados anteriormente como relevantes para a mídia do automobilismo para criar uma comparação entre eles, um subconjunto do nosso espaço de estados. Seja* $G = (S, E)$ *um grafo tal que* $S \subset V$ *definido por*

$$S = \{a, b, c, d, e\}$$

tal que

- $a =$ *Michael Schumacher;*
- $b =$ *Fernando Alonso;*
- $c =$ *Sebastian Vettel;*
- $d =$ *Lewis Hamilton;*
- $e =$ *Max Verstappen.*

Entre esses pilotos, somente Verstappen e Michael Schumacher não correram juntos, portanto, sua matriz de adjacência é

$$A = \begin{bmatrix} 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 & 0 \end{bmatrix}.$$

e sua respectiva representação pictórica do grafo utilizado é a seguinte:

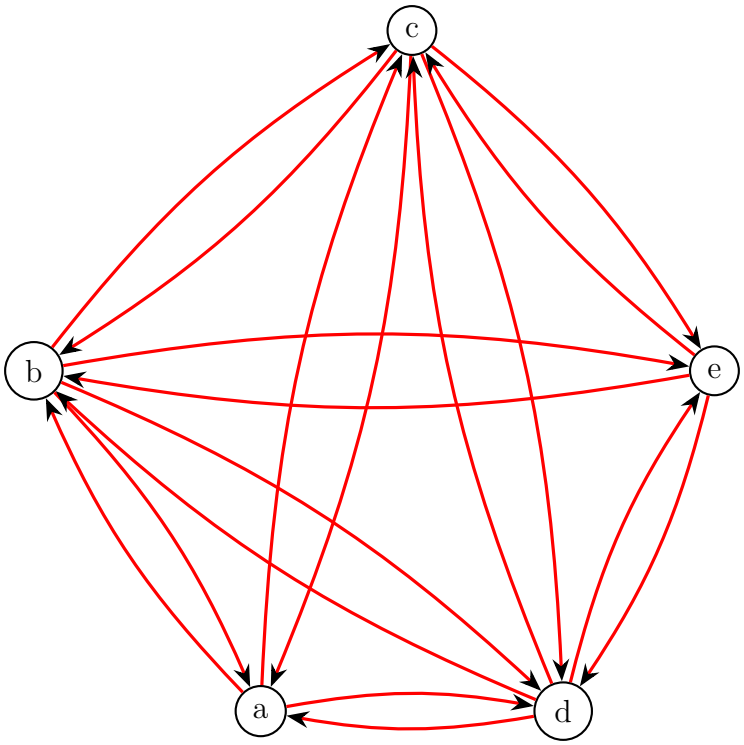

Figura 4.1: Representação pictórica do exemplo final.

Vamos, agora, observar os dados totais de cada piloto citado:

Tabela 4.1: Estatísticas das carreiras dos 5 pilotos escolhidos para o exemplo final.

Piloto	Corridas	Títulos	Vitórias	Pódios	Pontos Padronizados
Schumacher	308	7	91	155	3910
Alonso	358	2	32	98	2922
Vettel	300	4	53	122	3325
Hamilton	310	7	103	191	4796
Verstappen	163	2	35	77	2011

Para construirmos a função f para cada par de piloto, é necessário, além de limitar as estatísticas para quando cada par corria junto, determinar algum método para construir os βs. Um método simples que será utilizado é o seguinte: Todos os pilotos do exemplo correram a partir da década de 1990. Os campeões a partir do ano de 1990 em média tiveram as seguintes estatísticas:

- *8 vitórias;*
- *20 pódios;*
- *317 pontos padronizados.*

Dessa forma, assumimos que um título tem o mesmo peso de 8 vitórias, 20 pódios e 317 pontos padronizados. Portanto, definimos os βs de f como

$$f(a,b) = 317x_1^{(a,b)} + 39.625x_2^{(a,b)} + 15.85x_3^{(a,b)} + 1x_4^{(a,b)}.$$

Após os cálculos utilizando as estatísticas somente de quando o par de pilotos correu simultaneamente, obtemos os seguintes valores, que são disponibilizados em forma de matriz. Chamamos tal matriz de F e o elemento da linha i e coluna j representa $f(i,j)$.

$$F = \begin{bmatrix} 0 & 4844,25 & 213,85 & 213,85 & 0 \\ 4420,50 & 0 & 3262,50 & 3628,48 & 310,85 \\ 3310,58 & 8044,38 & 0 & 8626,82 & 2939,35 \\ 1408,95 & 11248,30 & 13780,40 & 0 & 9215,60 \\ 0 & 4216,20 & 5160,85 & 5241,32 & 0 \end{bmatrix}.$$

Com os valores de f para cada par, é possível calcular o valor dos respectivos ω. Seja Ω uma matriz tal que o elemento da linha i e coluna j representa $\omega(i \to j)$.

$$\Omega = \begin{bmatrix} 0 & 0,9125 & 15,4808 & 6,5885 & 0 \\ 1,0959 & 0 & 2,4657 & 3,1001 & 13,5634 \\ 0,0646 & 0,4056 & 0 & 1,5974 & 1,7558 \\ 0,1518 & 0,3226 & 0,6260 & 0 & 0,5687 \\ 0 & 0,0737 & 0,5695 & 1,7582 & 0 \end{bmatrix}.$$

Para encontrarmos a matriz de transição da cadeia que caracteriza a caminhada pelo grafo G, basta normalizarmos cada linha de forma que a soma de seus valores resultem em 1. Seja P a matriz de transição para a cadeia que queremos construir.

$$P = \begin{bmatrix} 0 & 0,0397 & 0,6736 & 0,2867 & 0 \\ 0,0543 & 0 & 0,1219 & 0,1533 & 0,6706 \\ 0,0169 & 0,1061 & 0 & 0,4178 & 0,4592 \\ 0,0909 & 0,1933 & 0,3751 & 0 & 0,3407 \\ 0 & 0,0307 & 0,2372 & 0,7321 & 0 \end{bmatrix}.$$

Como estamos trabalhando com uma matriz pequena, é possível encontrar a distribuição estacionária via multiplicação da matriz de transição e ranquear os 5 pilotos. A

distribuição estacionária π encontrada para este exemplo é

$$\pi = (0,04; 0,10; 0,23; 0,34; 0,29).$$

Desta forma, o ranking encontrado é:

1. *Lewis Hamilton;*
2. *Max Verstappen;*
3. *Sebastian Vettel;*
4. *Fernando Alonso;*
5. *Michael Schumacher.*

É interessante analisar os resultados encontrados na comparação entre tais pilotos. Por mais que Schumacher tenha o mesmo número de títulos que Hamilton, ele está em último enquanto o inglês está em primeiro. Isso se dá pelo fato de Schumacher ter ganhado seus títulos enquanto a maioria destes pilotos não corriam ainda (Alonso é a exceção). Assim, todos os outros demonstraram superioridade em cima de Schumacher enquanto corriam juntos e estão acima deles no ranking. O contrário pode ser dito a Verstappen: Enquanto ele corria com esses pilotos, era superior a eles, pois, retirando Hamilton, todos já tinham deixado de ativamente disputar por títulos e vitórias. Hamilton foi o único que mostrou superioridade e competição a todos enquanto corriam juntos, assim, faz sentido ele estar em primeiro.

No próximo capítulo é utilizado o método da multiplicação matricial para o cálculo dos diferentes *rankings*. Para casos em que computacionalmente não seja possível tal multiplicação, é necessária a utilização do método de simulação de cadeias de Markov. Para melhor entendimento sobre o passo a passo envolvendo a simulação de processos estocásticos, veja o apêndice C.

Capítulo 5

Resultados e Discussão

Neste capítulo, a partir do exemplo que foi construído no Capítulo 4, extrapolamos o ranqueamento que antes se limitava aos 5 pilotos para todos os pilotos que já competiram na Fórmula 1. Após a construção de um *ranking* inicial, diferentes abordagens da metodologia proposta nos capítulos anteriores foram tomadas afim de entender em mais detalhes o processo de ranqueamento.

5.1 O banco de dados

A análise estatística deste trabalho foi desenvolvida a partir de um banco de dados gratuito e constantemente atualizado disponível no repositório Kaggle, cujo *link* para *download* é: Kaggle - Formula 1 Data Base.

Esse banco de dados é constituído de 14 tabelas. Cada tabela armazena informações de diferentes naturezas. Nesse trabalho, as tabelas utilizadas no processo de análise estatística são:

- **Circuits**: Nome e país de todos os circuitos que já sediaram um grande prêmio da Fórmula 1;
- **Drivers**: Nome e Nacionalidade de todos os pilotos que já participaram de uma corrida;
- **Races**: Nome e data de cada corrida da história;
- **Seasons**: Ano e referência (Wikipedia) de cada temporada da história;
- **Driver Standings**: Resultado e pontuação de cada corrida da história;

- **Qualifying**: Resultado das classificatórias pré-corrida de cada corrida da história;
- **Results**: Resultado, pontuação e *status* final de cada piloto em cada corrida da história (mais detalhamento da tabela "Driver Standings");
- **Sprint Results**: Resultado e pontuação de cada corrida *sprint* da história.

Como temos à disposição, informações detalhadas a nível corrida a corrida, conseguimos verificar quando que dois pilotos correram juntos e, assim, aplicar a metodologia de passeios aleatórios sobre grafos para identificar o melhor piloto da Fórmula 1.

É importante ressaltar que, ao manipular os dados, percebemos que faltam informações das classificações dos anos 80 na tabela "*Qualifying*". Houve uma busca por tais informações em outros lugares, porém não foi encontrado em nenhum lugar estatísticas que fossem confiáveis e gratuitas. Portanto, por mais que acreditamos ser interessante levar em consideração as informações da classificação dos anos 80 no processo de análise estatística, decidimos por não considerá-las.

5.2 Ranqueamento via multiplicação de matrizes

O objetivo é construir o mesmo ranqueamento feito no Exemplo 4.1, seguindo o passo a passo descrito na Subseção 4.1, considerando todos os 857 pilotos que já correram na Fórmula 1. Para isso, é necessário superar os seguintes problemas que não se mostram presentes no exemplo devido ao baixo número de pilotos considerado, mas que se tornam evidentes ao aumentar o tamanho da amostra utilizada:

- Para quantificar o desempenho do piloto a em relação ao piloto b, no período em que eles correm juntos, utilizamos a função

$$f_1(a,b) = 317x_1^{(a,b)} + 39.625x_2^{(a,b)} + 15.85x_3^{(a,b)} + 1x_4^{(a,b)}$$

e, em seguida, construímos a matriz Ω cujas entradas são dadas por

$$\frac{f_1(b,a)}{f_1(a,b)}.$$

Dessa forma, quando $f_1(a,b) = 0$, a matriz Ω não fica bem definida. Para que este problema seja evitado, foi removido da amostra analisada todos os pilotos que nunca

pontuaram na carreira (com o sistema de pontos padronizado), ou seja, que nunca chegaram entre os 10 primeiros numa corrida. Dessa forma, $f_1(a, b) > 0$ para toda dupla de pilotos (a, b). Note que essa exclusão não possui um efeito significativo sobre *ranking* dos pilotos, pois o objetivo principal é ranquear os melhores pilotos e, um piloto que nunca conseguiu chegar entre os melhores em uma única corrida sequer pode ser considerado como um dos potenciais melhores pilotos da história. Após a exclusão, o tamanho da amostra é de 525 pilotos.

Mesmo com a restrição acima, como a comparação dois a dois é feita restringindo as estatísticas somente para as corridas em que ambos pilotos correram, ainda é possível haver casos em que o piloto não pontuou em nenhuma das corridas selecionadas para a comparação e, por consequência, a função f_1 possui valor igual a zero.

Para que estes casos também sejam evitados, foi dado um valor mínimo para a função $f_1(a, b) = 1$, como se o piloto tivesse pontuado uma única vez na carreira enquanto corria contra o piloto em comparação. Dessa forma, a divisão $\frac{f_1(b,a)}{f_1(a,b)}$ para a construção da matriz Ω se torna factível, pois sempre será positiva e, além disso, seus resultados para cada dupla continua de encontro com o objetivo da função, ou seja, quando $\omega(a, b) > 1$, o piloto a é considerado melhor que o piloto b e o contrário para quando $0 < \omega(a, b) < 1$.

5.2.1 Ranking com todos os pilotos que já pontuaram na história

Seja $G_1 = (V_1, E_1)$ um grafo, em que V_1 representa o conjunto de todos os pilotos que já pontuaram ao menos uma vez na história da Formula 1 e E_1 todos os elos que ligam tais pilotos que já correram juntos. Utilizando os códigos anexados no Apêndice D, foi encontrado os primeiros 20 para o seguinte ranking:

Tabela 5.1: Primeiros 20 pilotos do ranking utilizando a amostra de todos os pilotos que já pontuaram ao menos uma vez na Fórmula 1.

Rank	Piloto	Corridas	Títulos	Vitórias	Podiums	Pontos Std.
1	Alberto Ascari	36	2	13	17	443
2	Juan Manuel Fangio	58	5	24	35	876
3	Nino Farina	37	1	5	20	452
4	Dorino Serafini	1	0	0	1	18
5	Karl Kling	12	0	0	2	76
6	Oscar Gálvez	1	0	0	0	10
7	Cesare Perdisa	9	0	0	2	65
8	Jack Brabham	129	3	14	31	955
9	Eugenio Castellotti	17	0	0	3	86
10	Mike Hawthorn	48	1	3	18	469
11	Denny Hulme	112	1	8	33	951
12	José Froilán Gonzalez	29	0	2	15	302
13	Jim Clark	73	2	25	32	847
14	Jackie Stewart	100	3	27	43	1118
15	Alfonso de Portago	6	0	0	1	31
16	Sterling Moss	73	0	16	24	632
17	Jochen Rindt	62	1	6	13	367
18	John Barber	1	0	0	0	4
19	Luigi Fagioli	8	0	1	6	112
20	Nello Pagani	1	0	0	0	6

Ao analisar o ranking da Tabela 5.1, é possível notar que, ao considerar a função quantificadora f_1 e utilizar como amostra todos os pilotos que já pontuaram ao menos uma vez na história, o ranking é enviesado para pilotos que tiveram poucas corridas e correram nos anos 50 em algum momento. Todos os pilotos que estão nas primeiras 20 posições correram nos anos 50 em algum momento de sua carreira. Pilotos contemporâneos de destaque como Lewis Hamilton e Michael Schumacher, que naturalmente deveriam estar entre os mais bem colocados, estão nas posições 97 e 111, respectivamente, em um ranking de 525 pilotos.

Além disso, nota-se que pilotos que correram poucas corridas e nessas poucas, pontuaram bem, estão supervalorizados. Isso fica evidente ao perceber que, dos 20 primeiros, 7 possuem menos de 10 corridas. Tais pilotos tiveram ótimos resultados nas corridas participadas e, portanto, possuem um valor de comparação alto na grande maioria das comparações entre os pilotos que correram juntos. Com isso, tais pilotos acabam ficando bem ranqueados. Esse efeito é potencializado com o viés para os pilotos antigos, mencionado no parágrafo acima.

Como o objetivo é ranquear os melhores pilotos da F1, podemos criar um método de exclusão cujo impacto no objetivo de encontrar o melhor piloto é mínimo e não implicarão significativamente na comparação com outros companheiros de corrida. Além disso, o modelo ficará mais robusto para obtenção dos resultados.

5.2.2 Ranking com todos os pilotos que já venceram uma corrida ou já correram mais vezes que a média (30 corridas).

Seja $G_2 = (V_2, E_2)$ um grafo, em que V_2 representa o conjunto de todos os pilotos que já venceram uma corrida ao menos uma vez na história da Fórmula 1 ou correram mais que 30 corridas, o número médio de corridas que um piloto de Fórmula 1 participa, e E_2 todos os elos que ligam tais pilotos que já correram juntos.

Ao utilizar essa restrição, estamos retirando pilotos que, nas poucas corridas que participaram, em nenhuma se destacaram como os melhores dentre os que estavam correndo. Dessa forma, não estamos retirando de nossa comparação pilotos com extrema relevância para o ranqueamento. Para não excluirmos também pilotos com muitas corridas que nunca ganharam uma corrida, foi colocado a condição de que aqueles pilotos que já correram mais que a média de corridas de um piloto na Fórmula 1 (30 corridas) são incluídos na análise.

Para este grafo, os 20 primeiros colocados do ranking com 232 pilotos mostrado na Tabela 5.2 são:

Tabela 5.2: Primeiros 20 pilotos do ranking utilizando a amostra de todos os pilotos que já venceram ou correram mais de 30 corridas na Formula 1.

Rank	Piloto	Corridas	Títulos	Vitórias	Podiums	Pontos Std.
1	Juan Manuel Fangio	58	5	24	35	876
2	Alberto Ascari	36	2	13	17	443
3	Jim Clark	73	2	25	32	847
4	Nino Farina	37	1	5	20	452
5	Luigi Fagioli	8	0	1	6	112
6	Jackie Stewart	100	3	27	43	1118
7	Jack Brabham	129	3	14	31	955
8	Mike Hawthorn	48	1	3	18	469
9	Denny Hulme	112	1	8	33	951
10	Jochen Rindt	62	1	6	13	367
11	Jackie Oliver	51	0	0	2	101
12	Sterling Moss	73	0	16	24	632
13	José Froilán Gonzalez	29	0	2	15	302
14	Ludovico Scarfiotti	13	0	1	1	67
15	Phil Hill	52	1	3	16	367
16	Giancarlo Baghetti	26	0	1	1	65
17	Emerson Fittipaldi	149	2	14	35	994
18	Ronnie Peterson	123	0	10	26	733
19	Bruce McLaren	103	0	4	27	769
20	Piero Taruffi	18	0	1	5	156

Primeiramente, nota-se que ao utilizar outra amostra que a ordem da sequência entre os pilotos que estão nas duas amostras teve pouquíssimas alterações.

Também percebe-se que o viés perante os pilotos antigos continua, todos os pilotos dentre os top 20 são da era pré-anos 80. Comparando mais uma vez com Lewis e Michael, ambos estão nas posições 61 e 54, respectivamente, em um ranking de 230 pilotos.

Apesar do problema de pilotos com poucas corridas ter sido evitado na sua grande maioria, pilotos como Luigi Fagioli ainda estão em uma colocação muito acima do esperado. Pilotos com poucas corridas que corriam bem contra pilotos multi-campeões

acabam ficando bem colocados, o que explica pilotos como o número de corridas abaixo da média e resultados pouco expressivo entre os melhores. No caso de Luigi Fagioli, ele era contemporâneo e disputava o campeonato contra Juan Manuel Fangio e Alberto Ascari, os dois primeiros do ranking da Tabela 5.2.

Para evitar o novo problema citado, foi feito uma nova amostra.

5.2.3 Ranking com todos os pilotos que já venceram uma corrida e já correram mais vezes que a média (30 corridas)

Seja $G_3 = (V_3, E_3)$ um grafo, em que V_3 representa o conjunto de todos os pilotos que já venceram uma corrida ao menos uma vez na história da Formula 1 e correram mais que 30 corridas, o número médio de corridas que um piloto de Formula 1 participa, e E_3 todos os elos que ligam tais pilotos que já correram juntos.

Com este corte, temos um ranking somente com pilotos que já venceram no mínimo uma corrida e que tiveram um número expressivo de corridas para que faça sentido as comparações. Ao fazer este corte, estamos tirando pilotos com uma média de 1,3 podiums e 114 pontos padronizados na carreira, com pouca expressão para a história. Com isso, o ranqueamento continua não sendo prejudicado.

Os primeiros 20 colocados para o novo ranking com 96 pilotos são mostrados na tabela 5.3.

Tabela 5.3: Primeiros 20 pilotos do ranking utilizando a amostra de todos os pilotos que já venceram e correram mais de 30 corridas na Formula 1.

Rank	Piloto	Corridas	Títulos	Vitórias	Podiums	Pontos Std.
1	Juan Manuel Fangio	58	5	24	35	876
2	Alberto Ascari	36	2	13	17	443
3	Nino Farina	37	1	5	20	452
4	Mike Hawthorn	48	1	3	18	469
5	Sterling Moss	73	0	16	24	632
6	Jim Clark	73	2	25	32	847
7	Jochen Rindt	62	1	6	13	367
8	Phil Hill	52	1	3	16	367
9	Denny Hulme	112	1	8	33	951
10	Jackie Stewart	100	3	27	43	1118
11	Jack Brabham	129	3	14	31	955
12	Peter Collins	37	0	3	9	227
13	Maurice Trintignant	87	0	2	10	394
14	Emerson Fittipaldi	149	2	14	35	994
15	Tony Brooks	41	0	6	10	278
16	Graham Hill	179	2	14	35	994
17	Richie Ginther	54	0	1	14	436
18	Ronnie Peterson	123	0	10	26	733
19	John Surtees	112	1	6	24	688
20	Clay Regazzoni	138	0	5	28	824

Com a alteração da amostra utilizada, o problema foi resolvido, porém o viés para pilotos anteriores aos anos 80 continua. É interessante notar que a sequência dos pilotos também pouco muda quando é alterado somente a amostra utilizada. Para tentar corrigir tal viés, diferentes funções quantificadoras podem ser utilizadas, a fim de ter um ranqueamento em que não seja tão evidente a separação entre décadas.

5.3 Outras formas de construir a função f.

Para a amostra V_3, foi utilizada as seguintes funções quantificadoras f visando a mescla entre os pilotos de diferentes décadas, sem deixar de utilizar os pesos para cada variável definidos no exemplo 4.1:

1. $f_2(a,b) = (4 \times 317)x_1^{(a,b)} + (3 \times 39.625)x_2^{(a,b)} + (2 \times 15.85)x_3^{(a,b)} + x_4^{(a,b)}$.
2. $f_3(a,b) = (317^4)x_1^{(a,b)} + (39.625^3)x_2^{(a,b)} + (15.85^2)x_3^{(a,b)} + x_4^{(a,b)}$.
3. $f_4(a,b) = 317^{x_1^{(a,b)}} + (39.625^3)x_2^{(a,b)} + (2 * 15.85)x_3^{(a,b)} + x_4^{(a,b)}/10$.

A ideia da função quantificadora f_1 é dar pesos iguais para as estatísticas médias que um campeão alcança em sua campanha ao título. Com esta concepção, os multicampeões da era moderna estão sendo subestimados, portanto, a ideia de f_2 é criar uma discriminação maior entre cada uma das estatísticas dando um peso maior para os títulos, vitórias e pódiums, respectivamente.

A partir de agora, todos os rankings serão calculados utilizando a amostra V_3.

Ao fazer o cálculo do ranqueamento utilizando f_2, percebe-se que o viés para pilotos antigos continua, com poucas alterações entre o ranking gerado pela função f_1. Os primeiros 10 lugares do novo ranking é o seguinte:

Tabela 5.4: Primeiros 10 pilotos do ranking utilizando a função f_2 e a amostra de todos os pilotos que já venceram e correram mais de 30 corridas na Formula 1.

Rank	Piloto	Corridas	Títulos	Vitórias	Podiums	Pontos Std.
1	Juan Manuel Fangio	58	5	24	35	876
2	Alberto Ascari	36	2	13	17	443
3	Nino Farina	37	1	5	20	452
4	Mike Hawthorn	48	1	3	18	469
5	Phil Hill	52	1	3	16	367
6	Sterling Moss	73	0	16	24	632
7	Jim Clark	73	2	25	32	847
8	Denny Hulme	112	1	8	33	951
9	Jochen Rindt	62	1	6	13	367
10	Jackie Stewart	100	3	27	43	1118

Dado o resultado do ranking utilizando f_2, a ideia para a construção de f_3 é utilizar uma discriminação ainda maior entre cada estatística. Em f_2 foi utilizado a multiplicação na construção dos *beta* para quantificar o peso de cada variável e, para f_3, foi utilizado a potenciação. Os 10 primeiros do ranking utilizando a função f_3 é o seguinte:

Tabela 5.5: Primeiros 10 pilotos do ranking utilizando a função f_3 e a amostra de todos os pilotos que já venceram e correram mais de 30 corridas na Formula 1.

Rank	Piloto	Corridas	Títulos	Vitórias	Podiums	Pontos Std.
1	Juan Manuel Fangio	58	5	24	35	876
2	Alberto Ascari	36	2	13	17	443
3	Mike Hawthorn	48	1	3	18	469
4	Nino Farina	37	1	5	20	452
5	Phil Hill	52	1	3	16	367
6	Peter Collins	37	0	3	9	227
7	Sterling Moss	73	0	16	24	632
8	Max Verstappen	163	2	35	77	2000
9	Lewis Hamilton	310	7	103	191	4796
10	Pierre Gasly	108	0	1	3	336

O resultado é positivo porém não o suficiente. O viés continua para pilotos antigos, porém com menor instensidade. Como podemos observar no ranking da Tabela 5.5, os 5 primeiros continuam em uma sequência semelhante, porém nomes recentes como Max Verstappen e Lewis Hamilton são aparições esperadas e que em rankings anteriores não estavam nem perto dos 20 primeiros.

Nota-se que pilotos com poucos ou nenhum título se configuram com frequência no topo dos rankings, o que mostra que a variável "Título"não está bem ponderada. Vários multicampeões que são esperados estarem nas primeiras colocações em um ranking coerente estão no meio do ranking. Além disso, pilotos com a variável "Pontos Std."alta, como Peter Collins e Pierre Gasly, estão ficando melhor ranqueados do que o esperado. Por conta disso, a ideia de discriminação foi levada ao extremo em f_4.

A ideia inicial para a construção da função f_4 era $f_4 = \beta_i^{(x_i)}$ para $i = 1, 2, 3, 4$, porém, com excessão da variável de títulos, cujo o maior valor encontrado é 7, todos os outros coeficientes, ao serem calculados, encontram valores imensuráveis para um computador em algum momento, portanto, a contrução foi feita continuando com a multiplicação para

todos os β, com excessão do β da variável "Título". Utilizando a função f_4, os 10 primeiros do ranking é o seguinte:

Tabela 5.6: Primeiros 10 pilotos do ranking utilizando a função f_4 e a amostra de todos os pilotos que já venceram e correram mais de 30 corridas na Formula 1.

Rank	Piloto	Corridas	Títulos	Vitórias	Podiums	Pontos Std.
1	Lewis Hamilton	310	7	103	191	4796
2	Sebastian Vettel	300	4	53	122	3325
3	Mark Webber	217	0	9	42	1381
4	Michael Schumacher	308	7	91	155	3910
5	Rubens Barrichello	326	0	11	68	1906
6	Juan Pablo Montoya	95	0	7	30	828
7	Juan Manuel Fangio	58	5	24	35	876
8	Heikki Kovalainen	112	0	1	4	285
9	Phil Hill	52	1	3	16	367
10	Pastor Maldonado	96	0	1	1	76

Já em f_4, a discriminação entre as variáveis foram tantas que, como consequência, os multicampeões, que tendem a sempre ser muito acessados pelos pilotos que correram juntos, influenciaram muito positivamente em companheiros de equipes que corriam próximo a ele enquanto estavam com o mesmo carro na temporada, casos como Mark Webber, companheiro de equipe de Sebastian Vettel entre 2009 e 2013 e Rubens Barrichello, companheiro de Michael Schumacher entre 2000 e 2005. Por conta disso, a função final que foi utilizada leva em consideração algumas discriminações ao mesmo tempo que não é tão radical quanto a f_4. A função final f_5 utilizada é:

$$f_5(a,b) = 317^{x_1^{(a,b)}} + 39.625x_2^{(a,b)} + 15.85x_3^{(a,b)} + 1x_4^{(a,b)}.$$

Utilizando a função quantificadora f_5, os 10 primeiros lugares do ranking construído é o seguinte:

Tabela 5.7: Primeiros 10 pilotos do ranking utilizando a função f_5 e a amostra de todos os pilotos que já venceram e correram mais de 30 corridas na Formula 1.

Rank	Piloto	Corridas	Títulos	Vitórias	Podiums	Pontos Std.
1	Lewis Hamilton	310	7	103	191	4796
2	Sebastian Vettel	300	4	53	122	3325
3	Mark Webber	217	0	9	42	1381
4	Michael Schumacher	308	7	91	155	3910
5	Rubens Barrichello	326	0	11	68	1906
6	Heikki Kovalainen	112	0	1	4	285
7	Jarno Trulli	256	0	1	11	817
8	Juan Manuel Fangio	58	5	24	35	876
9	Jack Brabham	129	3	14	31	955
10	Phil Hill	52	1	3	16	367

Utilizando f_5, o problema dos pilotos companheiros de multicampeões que correram bem enquanto eles eram campeões permanece com menor intensidade. A efeito de curiosidade e retirando da comparação tais pilotos para um ranqueamento final, foi feito mais um corte na amostra, desta vez foi utilizado somente pilotos que já foram campeões ao menos uma vez na carreira.

5.3.1 Ranking com todos os pilotos que já foram campeões utilizando a função 5

Seja $G_4 = (V_4, E_4)$ um grafo cujo $V_4 \subset V$ e $E4 \subset E$ onde V_4 representa o conjunto de todos os pilotos que já campeões de ao menos um campeonato na história da Formula 1, e E_2 todos os elos que ligam tais pilotos que já correram juntos.

Seja também

$$f_5(a,b) = 317^{x_1^{(a,b)}} + 39.625x_2^{(a,b)} + 15.85x_3^{(a,b)} + 1x_4^{(a,b)}.$$

a função utilizada para o cálculo da matriz F.

Os 20 primeiros colocados para o novo ranking com 35 pilotos são:

Tabela 5.8: Primeiros 20 pilotos do ranking utilizando a função f_5 e a amostra de todos os pilotos que já venceram um campeonato na Formula 1.

Rank	Piloto	Corridas	Títulos	Vitórias	Podiums	Pontos Std.
1	Lewis Hamilton	310	7	103	191	4796
2	Sebastian Vettel	300	4	53	122	3325
3	Michael Schumacher	308	7	91	155	3910
4	Juan Manuel Fangio	58	5	24	35	876
5	Jack Brabham	129	3	14	31	955
6	Phil Hill	52	1	3	16	367
7	Jackie Stewart	100	3	27	43	1118
8	Graham Hill	179	2	14	36	1075
9	Jim Clark	73	2	25	32	847
10	Jochen Rindt	62	1	6	13	367
11	Denny Hulme	112	1	8	33	951
12	Nico Rosberg	206	1	23	57	1759
13	Emerson Fittipaldi	149	2	14	35	994
14	James Hunt	93	1	10	23	630
15	Niki Lauda	174	3	25	54	1348
16	John Surtees	112	1	6	24	688
17	Alain Prost	202	4	51	106	2486
18	Ayrton Senna	162	3	41	80	1885
19	Nelson Piquet	207	3	23	60	1692
20	Jody Scheckter	113	1	10	33	899

Por mais que sejam alteradas a amostra e a função utilizadas, podemos perceber que ainda temos alguns conglomerados entre pilotos de mesma época, o que evidencia que o tempo cujo piloto corria ainda está afetando o ranqueamento, o que será mais discutido nas considerações finais.

Capítulo 6

Considerações finais

Relembrando o objetivo inicial da pesquisa, buscamos identificar, a partir de uma regra de decisão pré-estabelecida, o melhor piloto de Fórmula 1 da história. Ao longo da apresentação dos resultados e da discussão sobre a metodologia a ser utilizada, vários rankings foram construídos a partir de diferentes amostras e funções. A partir dessas variações, podemos tirar algumas conclusões sobre o processo.

Em uma análise mais a fundo, notamos que os pilotos contemporâneos, ou seja, aqueles que correram por muito tempo juntos, tendem a ficar conglomerados no ranking e a manter suas posições relativas entre si, mesmo com mudanças nas amostras utilizadas. No entanto, suas ordens relativas se alteram quando há mudanças na função utilizada. Quando utilizamos diferentes amostras, as posições de cada conglomerado mudam, pois o peso dado a cada década varia. Independente do ranking construído, ele será bem fundamentado com uma amostra representativa dos indivíduos avaliados e uma função que capture adequadamente o desempenho de cada piloto.

Um dos principais objetivos era eliminar o efeito temporal na análise, mas o resultado obtido não foi o esperado. Como mencionado anteriormente, os pilotos contemporâneos continuaram sendo agrupados, e ao alterar a amostra utilizada no ranking, percebemos que as posições relativas se alteravam entre décadas, mas não o agrupamento. Após diversas análises, notamos que as décadas de 1950 e 2000, anos sem supremacias individuais no esporte, sempre figuram no topo dos rankings. Já as décadas de 2010 e 1980, anos com maiores supremacias individuais, tendem a ficar abaixo no ranking. Essa análise sugere que grandes supremacias são construídas sobre pilotos menos competitivos. Assim, pilotos como Senna, Prost e Verstappen, que geralmente aparecem em posições altas na maioria dos rankings, não estão bem posicionados.

Foi mencionado ao longo do texto a possível necessidade de utilizar métodos computacionais de simulação de cadeia, caso a multiplicação matricial se tornasse um empecilho devido ao grande número de amostras. Em termos de tempo de cálculo dos rankings, a maior parte é gasta no cálculo da matriz F. No entanto, ao fazer as multiplicações de P em busca de Π, o resultado é obtido rapidamente. Por isso, a simulação computacional das cadeias não foi implementada, mas sua definição e o passo a passo de como utilizá-la estão descritos no apêndice C.

O ranqueamento construído pode ser aprimorado e utilizado para outras aplicações. O fato de pilotos contemporâneos ainda ficarem agrupados indica que o efeito temporal ainda influencia o resultado final, e análises adicionais podem ser feitas para mitigar esse efeito. Além disso, esse ranqueamento pode ser adaptado para diferentes esportes, com ajustes na amostra e na função f, para que ambos façam sentido. Bons exemplos de utilização do ranqueamento em outros esportes incluem a comparação de tenistas com base em suas estatísticas e pontuações na ATP ou WTA, ou a comparação de quarterbacks da NFL com base em suas estatísticas de confronto direto, especialmente a estatística de QBR, amplamente usada para medir a produtividade do atleta na partida.

Referências Bibliográficas

Barlow, J. (2022). Here are the 10 best ever formula 1 drivers. https://www.topgear.com/car-news/formula-one/here-are-10-best-ever-formula-1-drivers.

Bell, A., Smith, J., Sabel, C. E. e Jones, K. (2016). Formula for success: multilevel modelling of formula one driver and constructor performance, 1950–2014. *Journal of Quantitative Analysis in Sports*, **12**(2), 99–112.

Brémaud, P. (2013). *Markov chains: Gibbs fields, Monte Carlo simulation, and queues*, volume 31. Springer Science & Business Media.

Eichenberger, R., Stadelmann, D. *et al.* (2009). Who is the best formula 1 driver? an economic approach to evaluating talent. *Economic Analysis and Policy*, **39**(3), 389.

Ferrari, P. A. e Galves, A. (1997). Acoplamento e processos estocásticos.

GP, R. (2022). Quem são os melhores pilotos da história da fórmula 1. https://www.grandepremio.com.br/f1/noticias/os-melhores-pilotos-de-f1/.

Häggström, O. *et al.* (2002). *Finite Markov chains and algorithmic applications*, volume 52. Cambridge University Press.

Harden, O. (2015). The mclaren mp4/4 and 5 of formula 1's most dominant cars. https://bleacherreport.com/articles/2447060-the-mclaren-mp44-and-5-of-formula-1s-most-dominant-cars.

Jeffries, T. (2023). The 10 best formula 1 drivers ever: Hamilton, schumacher more. https://www.autosport.com/f1/news/whos-the-best-formula-1-driver-schumacher-hamilton-senna-more-4983210/4983210/.

Leathers, S. (2022). James hunt and niki lauda's unlikely friendship despite infamous rivalry. https://www.express.co.uk/sport/f1-autosport/1635619/james-hunt-niki-lauda-friendship-rivalry-formula-one-racing-spt.

Levin, D. A. e Peres, Y. (2017). *Markov chains and mixing times*, volume 107. American Mathematical Soc.

Lopes, R. (2022). Senna x prost: a maior rivalidade da história do esporte mundial. https://ge.globo.com/motor/formula-1/blogs/voando-baixo/post/2020/04/03/senna-x-prost-a-maior-rivalidade-da-historia-do-esporte-mundial.ghtml.

Manishin, G. B. (2018a). After tamburello. http://www.f1-grandprix.com/?page_id=1796.

Manishin, G. B. (2018b). The british era. http://www.f1-grandprix.com/?page_id=923.

Newman, M. (2018). *Networks*. Oxford University Press.

Phillips, A. J. (2014). Uncovering formula one driver performances from 1950 to 2013 by adjusting for team and competition effects. *Journal of Quantitative Analysis in Sports*, **10**(2), 261–278.

Piquero, A. R., Piquero, N. L. e Han, S. (2021). Identifying the most successful formula 1 drivers in the turbo era. *The Open Sports Sciences Journal*, **14**(1).

Reynolds, N. (2022). What are the best f1 tracks for racing? https://lastwordonsports.com/motorsports/2022/07/09/what-are-the-best-f1-tracks-for-racing/.

Rossini, M. C. (2023). Os maiores pilotos de fórmula 1, segundo o sistema de pontuação atual. https://super.abril.com.br/sociedade/os-maiores-pilotos-de-formula-1-segundo-o-sistema-de-pontuacao-atual/.

Straw, E. (2017). The ferrari that started the schumacher f1-dominance era. https://www.autosport.com/f1/news/the-ferrari-that-started-the-schumacher-f1-dominance-era-4988679/4988679/.

Apêndice A

Prova da Unicidade da Distribuição Estacionária

Para provarmos o teorema 3.33, precisamos antes da definição de alguns conceitos, como e de funções harmônicas e cadeias reversas.

Definição A.1 *Chamamos uma função $h : V \to \mathbb{R}$ de* ***função harmônica*** *quando*

$$h(a) = \sum_{b \in V} P(a,b)\, h(b).$$

Uma função é harmônica em V se for harmônica em todo estado $a \in V$. Se P é uma matriz e h é vista como vetor coluna, então a função que é harmônica em todos os estados de V satisfaz a equação $Ph = h$. Assim, temos o seguinte resultado, que será usado na a prova do teorema.

Proposição A.2 *Seja uma cadeia de Markov $\boldsymbol{X} := \{X_n : n \in \mathbb{N}\}$ irredutível a tempo discreto definida em um espaço de probabilidade $(\Omega, \mathcal{F}, \mathbb{P})$ e assumindo valores no espaço de estados V com matriz de transição P. Então, a função h harmônica em qualquer ponto de V como definido acima é constante.*

Definição A.3 *A* ***cadeia reversa*** *de uma cadeia de Markov $(X_0, X_1, ...)$ com distribuição estacionária π pode ser visto como outra cadeia de Markov com matriz de transição*

$$\hat{P}(a,b) := \frac{\pi(b)\, P(b,a)}{\pi(a)}.$$

A cadeia reversa é simplesmente a cadeia com os índices temporais revertidos. O

seguinte resultado também será utilizado na prova do teorema.

Proposição A.4 *Seja uma cadeia de Markov* $\boldsymbol{X} := \{X_n : n \in \mathbb{N}\}$ *irredutível a tempo discreto definida em um espaço de probabilidade* $(\Omega, \mathcal{F}, \mathbb{P})$ *e assumindo valores no espaço de estados* V *com matriz de transição* P *e distribuição estacionária* π*. Seja também* $\hat{\boldsymbol{X}}$ *a cadeia reversa de* $\boldsymbol{X}$ *com matriz de transição* $\hat{P}$*. Então* π *é a distribuição estacionária de* $\hat{P}$ *e para qualquer sequência* $(X_0, X_1, ..., X_n) \in V$*, temos*

$$P_\pi\{X_0 = a_0, X_1 = a_1, ..., X_n = a_n\} = \hat{P}_\pi\{X_0 = a_n, X_1 = a_{n-1}, ..., X_n = a_0\}.$$

Prova do teorema 3.33.

Dada duas distribuições estacionárias π_1 e π_2, defina

$$f(a) = \frac{\pi_1(a)}{\pi_2(a)}.$$

Queremos mostrar que $f(a)$ é harmônica, para isso,

$$\sum_{b \in V} P(a,b) f(b) = \sum_{b \in V} P(a,b) \frac{\pi_1(b)}{\pi_2(b)}.$$

Ao multiplicar o lado direito da equação por $\frac{\pi_2(a)}{\pi_2(a)}$, temos que

$$\sum_{b \in V} P(a,b) f(b) = \sum_{b \in V} P(a,b) \frac{\pi_2(a)}{\pi_2(b)} \frac{\pi_1(b)}{\pi_2(a)}.$$

Pela definição de cadeias reversas,

$$\sum_{b \in V} P(a,b) f(b) = \frac{1}{\pi_2(a)} \sum_{b \in V} \hat{P}(b,a) \pi_1(b).$$

Pela definição de medida estacionária e pela proposição A.4

$$\sum_{b \in V} P(a,b) f(b) = \frac{\pi_1(a)}{\pi_2(a)} = f(a)$$

o que nos mostra que a função f é harmônica.

A proposição A.2 nos mostra que, como a cadeia é irredutível, $f(a)$ é constante para

todo $a \in V$. Logo,

$$\frac{\pi_1(b)}{\pi_2(b)} = \frac{\pi_1(a)}{\pi_2(a)},\ \forall\, b \in V.$$

Podemos dizer também que $\frac{\pi_1(b)}{\pi_2(b)} = \frac{\pi_1(a)}{\pi_2(a)} = c$, ou seja,

$$\pi_1(b) = c\pi_2(b).$$

Por definição, temos que

$$\sum_{b\in V} \pi_1(b) = 1 \quad e \quad \sum_{b\in V} \pi_2(b) = 1.$$

Assuma $c \neq 1$.

$$1 = \sum_{b\in V} \pi_1(b) = \sum_{b\in V} c\pi_2(b) = c\sum_{b\in V} \pi_2(b) = c \neq 1.$$

Quando $c \neq 1$, temos um absurdo, Logo, $c = 1$ e $\pi_1(b) = \pi_2(b)\ \ \forall\ b \in V$ o que mostra que a distribuição estacionária é única.

Apêndice B

Prova do Teorema da Convergência

Antes de enunciarmos a prova do teorema, é necessário uma definição alternativa de distância de variação total.

Definição B.1 *Seja μ e ν duas distribuições de probabilidade em V. Então*

$$||\mu - \nu||_{TV} = \frac{1}{2}\sum_{a\in V}|\mu(a) - \nu(a)| \quad ou \quad ||\mu - \nu||_{TV} = \sum_{\mu(a)\geq\nu(a)}(\mu(a) - \nu(a)),\ \forall\ a \in V.$$

Prova do teorema 3.36.

Como P é irredutível e aperiódica, por definição, existe r tal que

$$P^r(a, y) > 0,\ \forall\ a, b \in V.$$

Seja Π a matriz com $|V|$ linhas na qual cada linha é o vetor π. Para $\delta > 0$ suficientemente pequeno, podemos assumir

$$P^r(a, b) \geq \delta\pi(b),\ \forall\ a, b \in V.$$

Seja $\Theta = 1 - \delta$. A equação

$$P^r = (1 - \Theta)\Pi + \Theta Q$$

define a matriz estocástica Q. Pela definição de matriz estocástica, para qualquer matriz estocástica M, temos que $M\Pi = \Pi$ e para qualquer matriz M (não necessariamente estocástica) tal que $\pi M = \pi$, temos que $\Pi M = \Pi$. Com isso, vamos mostrar que

$$P^{rk} = (1 - \Theta^k)\Pi + \Theta^k Q^k.$$

- Para $k = 1$, temos que já é válido.

- Assuma que a hipótese valha para $k = n$, ou seja, $P^{rn} = (1 - \Theta^n)\Pi + \Theta^n Q^n$.

- Para $k = (n + 1)$,

$$\begin{aligned} P^{r(n+1)} &= P^{rn}P^r = (1 - \Theta^n)\Pi + \Theta^n Q^n P^r = (1 - \Theta^n)\Pi P^r + \Theta^n Q^n P^r \\ &= (1 - \Theta^n)\Pi P^r + \Theta^n Q^n[(1 - \Theta)\Pi + \Theta Q] \\ &= (1 - \Theta^n)\Pi + (1 - \Theta)\Theta^n Q^n \Pi + \Theta^{n+1} Q^{n+1} \\ &= (1 - \Theta^n)\Pi + (1 - \Theta)\Theta^n \Pi + \Theta^{n+1} Q^{n+1} \\ &= [(1 - \Theta^n) + (1 - \Theta)\Theta^n]\Pi + \Theta^{n+1} Q^{n+1} \\ &= [1 - \Theta^n + \Theta^n - \Theta^{n+1}]\Pi + \Theta^{n+1} Q^{n+1} \\ &= [1 - \Theta^{n+1}]\Pi + \Theta^{n+1} Q^{n+1}. \end{aligned}$$

Ao multiplicar p^{rk} por P^j, temos que

$$\begin{aligned} P^{rk}P^j &= [(1 - \Theta^k)\Pi + \Theta^k Q^k]P^j = [(1 - \Theta^k)\Pi + \Theta^k Q^k]P^j = (1 - \Theta^k)\Pi P^j + \Theta^k Q^k P^j \\ &= (1 - \Theta^k)\Pi + \Theta^k Q^k P^j = \Pi - \Theta^k \Pi + \Theta^k Q^k P^j. \end{aligned}$$

Subtraindo Π de ambos os lados, concluímos que

$$\begin{aligned} P^{rk+j} - \Pi &= -\Theta^k[\Pi - Q^k P^j] \\ &= \Theta^k[Q^k P^j - \Pi]. \end{aligned}$$

Vamos focar nossa análise na primeira linha de P^{rk+j} chamada de a_0.

$$\begin{aligned} P^{rk+j}(a_0, .) - \Pi(a_0, .) &= \Theta^k[Q^k P^j(a_0, .) - \Pi(a_0, .)] \\ P^{rk+j}(a_0, .) - \pi &= \Theta^k[Q^k P^j(a_0, .) - \pi]. \end{aligned}$$

Ao utilizar um somatório sobre todos $a \in V$, a igualdade continua.

$$\sum_{a \in V} P^{rk+j}(a_0, a) - \pi(a) = \sum_{a \in V} \Theta^k (Q^k P^j(a_0, a) - \pi(a))$$
$$\frac{1}{2} \sum_{a \in V} P^{rk+j}(a_0, a) - \pi(a) = \frac{1}{2} \sum_{a \in V} \Theta^k (Q^k P^j(a_0, a) - \pi(a)).$$

Pela definição B.1, temos que

$$||P^{rk+j}(a_0, .) - \pi||_{TV} = \frac{1}{2} \Theta^k \sum_{a \in V} (Q^k P^j(a_0, a) - \pi(a)).$$

$Q^k P^j$ e Π são matrizes estocásticas, logo, a soma da subtração $(Q^k P^j(a_0, a) - \pi(a))$ é majorada por 1, e assim, temos a seguinte desigualdade

$$||P^{rk+j}(a_0, .) - \pi||_{TV} \leq \frac{1}{2} \Theta^k \leq \Theta^k.$$

Por fim, seja $t = (rk + j)$ ou $k = \frac{t-j}{r}$, com $j < r$ (onde r é fixo).

$$||P^t(a_0, .) - \pi||_{TV} \leq \Theta^{\frac{t-j}{r}} \leq \Theta^{\frac{t}{r}} \Theta^{\frac{-j}{r}} \leq (\Theta^{\frac{1}{r}})^t \Theta^{\frac{-j}{r}}.$$

Assuma $\Theta^{\frac{1}{r}} = \alpha$.

$$||P^t(a_0, .) - \pi||_{TV} \leq \alpha^t \Theta^{\frac{-j}{r}}.$$

Assuma $\Theta^{\frac{-j}{r}} = C$.

$$||P^t(a_0, .) - \pi||_{TV} \leq C\alpha^t.$$

Como vale para todos, vale para o maior entre as variações totais, portanto,

$$\max_{a \in V} ||P^t(a, .) - \pi||_{TV} \leq C\alpha^t.$$

Apêndice C

Método Alternativo para Simulação de Cadeias de Markov

A regra que utilizamos para decidir qual é o melhor piloto da história é baseada em passeios aleatórios através de uma matriz cujas probabilidades de transição entre dois vértices são funções das medidas de similaridade entre eles. Cada passeio aleatório iniciará sua trajetória em um vértice diferente do grafo e, após uma quantidade suficientemente grande de transições, vamos analisar quais foram os vértices mais visitados e, a partir deste número de visitas, identificaremos quais são os melhores pilotos de todos os tempos. Nesse sentido, apresentamos a seguir uma definição construtivista de cadeia de Markov, a qual será utilizada durante o processo de simulação.

Definição C.1 *Seja* $\boldsymbol{X} := \{X_n : n \in \mathbb{N}\}$ *um processo estocástico a tempo discreto definido em um espaço de probabilidade* $(\Omega, \mathcal{F}, \mathbb{P})$ *e assumindo valores no espaço de estados* V*. Tal processo é definido como uma* ***cadeia de Markov*** *se existir uma função* $\Phi : V \times [0,1] \to V$ *tal que, para todo* $n \geq 1$*,*

$$X_n = \Phi(X_{n-1}, U_n),$$

em que $U_1, U_2, \ldots$ *é uma sequência de variáveis aleatórias independentes entre si com distribuição uniforme no intervalo* $[0,1]$*.*

Exemplo C.2 *Considere a seguinte cadeia:*

Para este simples caso, $V = \{0,1\}$*. Podemos definir* $\Phi(x,u)$ *da seguinte forma:*

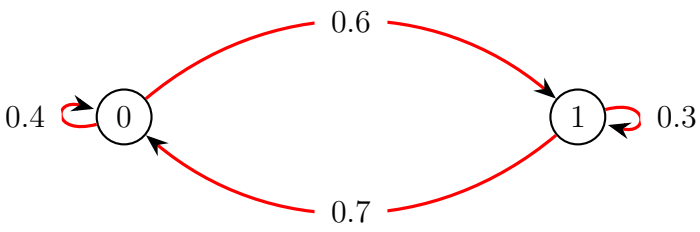

Figura C.1: Representação do grafo de transição de uma cadeia de dois estados com probabilidades de transição definidas.

$$\Phi(x, u) = \mathbf{1}\{u > h(x)\},$$

tal que $h(0) = 0.4$ *e* $h(1) = 0.3$.

Desta forma, em $X_{n-1} = 0$, *se* $U_n > 0.4$, *então* $X_n = 1$. *Caso, contrário,* $X_n = 0$. *Em outras palavras, dado que a cadeia está em* 0, *em* 40% *das vezes ela continua em* 0 *e* 60% *das vezes ela migra para o estado* 1.

O mesmo vale para quando $X_{n-1} = 1$. *Se* $U_n > 0.3$, *então* $X_n = 1$ *e caso contrário,* $X_n = 0$. *Neste caso, quando* $U_n > h(1)$, *a cadeia continua no mesmo estado e quando* $U_n < h(1)$, *a cadeia migra para o estado* 0.

Proposição C.3 *Seja* V *um conjunto finito ou enumerável. Dada uma matriz de transição* P *definida em* V *e um elemento* $a \in V$ *qualquer, é possível construir uma cadeia de Markov* $\mathbf{X}$, *tendo* P *como a matriz de transição e* a *como o estado inicial.*

Demonstração. *A prova da proposição pode ser encontrada em Ferrari e Galves (1997).*

□

A Definição C.1 e a Proposição C.3 são de suma importância, pois nos mostra que, dados um estado inicial e uma matriz de transição, precisamos somente definir uma função $\Psi(x, u)$ para descobrir os próximos estados e, assim, simular nossa cadeia. Ao invés de utilizar da multiplicação de matrizes para encontrar a distribuição estacionária, simulamos a cadeia um número grande o suficiente de vezes até que possamos assumir que a distribuição de visitas em cada estado seja a distribuição estacionária.

Para construir uma simulação de uma cadeia de Markov, precisamos então de duas funções: A **função de iniciação** e a **função de atualização**. A função de iniciação

$$\Psi : [0, 1] \to V$$

é uma função que parte de uma distribuição uniforme U_0 para um estado de V. A utilizamos para encontrar o valor de X_0, cuja construção se baseia em dividir a distribuição uniforme em $|V|$ partes iguais para a escolha aleatória de um estado inicial. Seja $V = \{v_1, v_2, ..., v_{|V|}\}$. A função de iniciação é dada por

$$\Psi(x) = v_i \quad \text{se } x \in \left[\frac{i-1}{|V|}, \frac{i}{|V|}\right]. \tag{C.4}$$

Agora que sabemos como encontrar X_0, precisamos de uma função que gera X_{n+1} a partir de X_n para qualquer $n \in \mathbb{N}$. A função de atualização $\Phi : V \times [0,1] \to V$ parte de um estado v e para um outro estado v' de acordo com o valor de uma distribuição uniforme U_n, tal que

$$\Phi(v_i, x) = \begin{cases} v_1 & \text{se } x \in [0, P(v_i, v_1)]\,; \\ \vdots & \\ v_j & \text{se } x \in \left[\sum_{k=1}^{j-1} P(v_i, v_k), \sum_{k=1}^{j} P(v_i, v_k)\right]; \\ \vdots & \\ v_{|V|} & \text{se } x \in \left[\sum_{k=1}^{j-1} P(v_i, v_k), 1\right]. \end{cases}$$

Assim, temos como simular todos os valores de qualquer cadeia que queremos, utilizando Ψ para definir o estado inicial e Φ para X_n.

Exemplo C.5 *Para o Exemplo 3.2 dos 5 pilotos, vamos construir as funções de iniciação e atualização que define o passeio no grafo que construímos. Para isso, considere a matriz P dada no exemplo 3.18:*

$$P = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0.125 & 0 & 0.375 & 0.5 & 0 \\ 0 & 0.33 & 0 & 0.67 & 0 \\ 0 & 0.2 & 0.3 & 0 & 0.5 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}.$$

Como $|V|$ neste caso é 5, a função de iniciação é dada por

$$\Psi(x) = \begin{cases} a & se\ x \in \left[0, \frac{1}{5}\right]; \\ b & se\ x \in \left[\frac{1}{5}, \frac{2}{5}\right]; \\ c & se\ x \in \left[\frac{2}{5}, \frac{3}{5}\right]; \\ d & se\ x \in \left[\frac{3}{5}, \frac{4}{5}\right]; \\ e & se\ x \in \left[\frac{4}{5}, 1\right]. \end{cases}$$

Já a função de atualização é dada para cada estado por:

- *Piloto A:*

$$\Psi(a, x) = \begin{cases} b & se\ x \in [0, 1]. \end{cases}$$

- *Piloto B:*

$$\Psi(b, x) = \begin{cases} a & se\ x \in [0, 0.125]; \\ c & se\ x \in [0.125, 0.5]; \\ d & se\ x \in [0.5, 1]. \end{cases}$$

- *Piloto C:*

$$\Psi(c, x) = \begin{cases} b & se\ x \in [0, 0.33]; \\ d & se\ x \in [0.33, 1]. \end{cases}$$

- *Piloto D:*

$$\Psi(d, x) = \begin{cases} b & se\ x \in [0, 0.2]; \\ c & se\ x \in [0.2, 0.5]; \\ e & se\ x \in [0.5, 1]. \end{cases}$$

- *Piloto E:*

$$\Psi(e, x) = \begin{cases} d & se\ x \in [0, 1]. \end{cases}$$

Com tais funções, podemos de forma iterativa construir passeios dentro da cadeia utilizando somente um estado inicial arbitrário, a matriz de transição e distribuições uniformes em $[0, 1]$ independentes.

Apêndice D

Códigos Utilizados

```
library(dplyr)

### TABLES ###

circuits = read.csv("circuits.csv")
circuits = circuits[,1:5]

driver_standings = read.csv("driver_standings.csv")
driver_standings = driver_standings[,c(1,2,3,4,5)]

drivers = read.csv("drivers.csv")
drivers = drivers[,c(1,2,4,5,6,8)]

qualifying = read.csv("qualifying.csv")
qualifying = qualifying[,c(1,2,3,4,6)]

races_table = read.csv("races.csv")
races_table = races_table[,1:6]

# Table for join with results #

results = read.csv("results.csv")
```

```
results = results[,c(1,2,3,4,9,10,15,18)]
colnames(results)[colnames(results) == 'points'] = 'race_points'

seasons = read.csv("seasons.csv")

sprint_results = read.csv("sprint_results.csv")
sprint_results = sprint_results[,c(1,2,3,4,9,10,16)]
colnames(sprint_results)[colnames(sprint_results) == 'points'] = 'sprint_points'

# Table to join sprint_points with results #

sprint_join_results = sprint_results[,c(2,3,6)]

status = read.csv("status.csv")

### Adding years and sprint_points to the results table ###

results = left_join(results, races_table %>% select(raceId, year))
results = left_join(results,
                    sprint_join_results,
                    by = c('raceId' = 'raceId', 'driverId' = 'driverId')
)
results$sprint_points[is.na(results$sprint_points)] = 0

results$points = results$sprint_points + results$race_points

### TITLES ###

driver_standings = left_join(driver_standings, races_table %>% select(raceId, year))

driverId = c()
year = c()
```

```
for(i in min(seasons[,1]):max(seasons[,1])) {
  temp = driver_standings[which(driver_standings$year == i),]
  campeao = temp[which(temp$raceId == max(temp$raceId) & temp$position == 1), 3]
  driverId = c(driverId, campeao)
  year = c(year,i)
}

champs = data.frame(driverId, year)

seasons = left_join(seasons, champs, by = c('year' = 'year'))
seasons = left_join(seasons,
                    drivers %>% select(driverId, driverRef, forename, surname)
)
seasons = seasons[-c(74),-c(2)]

count_champs = seasons %>% count(driverRef, sort = T)

drivers = left_join(drivers, count_champs %>% select(driverRef, n))
drivers[is.na(drivers)] = 0
colnames(drivers)[colnames(drivers) == 'n'] = 'titles'

### STANDARIZED POINTS ###

for (i in 1:length(results$resultId)) {
  if (results$positionOrder[i] == 1) {results$standarized_points[i] = 25}
  else if (results$positionOrder[i] == 2) {results$standarized_points[i] = 18}
  else if (results$positionOrder[i] == 3) {results$standarized_points[i] = 15}
  else if (results$positionOrder[i] == 4) {results$standarized_points[i] = 12}
  else if (results$positionOrder[i] == 5) {results$standarized_points[i] = 10}
  else if (results$positionOrder[i] == 6) {results$standarized_points[i] = 8}
  else if (results$positionOrder[i] == 7) {results$standarized_points[i] = 6}
  else if (results$positionOrder[i] == 8) {results$standarized_points[i] = 4}
  else if (results$positionOrder[i] == 9) {results$standarized_points[i] = 2}
```

```
  else if (results$positionOrder[i] == 10) {results$standarized_points[i] = 1}
  else {results$standarized_points[i] = 0}

  if (results$positionOrder[i]>0&results$positionOrder[i]<11&results$rank[i]==1){
    results$standarized_points[i] = results$standarized_points[i] + 1}
}

# See which competitor would be the champion #

std_champ = c()
std_champ_points = c()

for (i in year) {
  std_champ_year = results[which(results$year == i),]
  std_results_year = std_champ_year %>% group_by(driverId) %>%
    summarise(sum_points = sum(standarized_points), .groups = 'drop') %>%
    as.data.frame()
  std_results_year = std_results_year[order(std_results_year$sum_points,
                                            decreasing = TRUE),]
  std_champ = c(std_champ, std_results_year[1,1])
  std_champ_points = c(std_champ_points, std_results_year[1,2])
}

std_champs = data.frame(year, std_champ, std_champ_points)
std_seasons = left_join(seasons, std_champs, by = c('year' = 'year'))

std_seasons = left_join(std_seasons, drivers, by = c('std_champ' = 'driverId'))
std_seasons = std_seasons[,1:7]

colnames(std_seasons)[colnames(std_seasons) == 'driverId.x'] = 'driverId'
colnames(std_seasons)[colnames(std_seasons) == 'driverRef.x'] = 'driverRef'
colnames(std_seasons)[colnames(std_seasons) == 'forename.x'] = 'forename'
colnames(std_seasons)[colnames(std_seasons) == 'surname.x'] = 'surname'
```

```
colnames(std_seasons)[colnames(std_seasons) == 'driverRef.y'] = 'std_champ_ref'

### DATA FOR THE CHAMPION ###

champ_std_points = c()
champ_podiums = c()
champ_wins = c()
num_races = c()

for(i in 1:length(seasons[,1])) {
  subset_champ = results[
    which(results$year==seasons$year[i]&results$driverId==seasons$driverId[i]),]

  champ_std_points = c(champ_std_points,
                       sum(subset_champ$standarized_points +
                           subset_champ$sprint_points)
                      )

  count_champ_positions = subset_champ %>% count(positionOrder, sort = T)

  if(identical(count_champ_positions[
    which(count_champ_positions$positionOrder == 1), 2
  ], integer(0)) == FALSE){
        first_place = count_champ_positions[
            which(count_champ_positions$positionOrder == 1), 2
        ]
    } else {first_place = 0}

  if(identical(count_champ_positions[
    which(count_champ_positions$positionOrder == 2), 2
  ], integer(0)) == FALSE) {
    second_place = count_champ_positions[
        which(count_champ_positions$positionOrder == 2), 2
```

```
    ]
  } else {second_place = 0}

  if(identical(count_champ_positions[
    which(count_champ_positions$positionOrder == 3), 2
  ], integer(0)) == FALSE) {
    third_place = count_champ_positions[
        which(count_champ_positions$positionOrder == 3), 2
    ]
  } else {third_place = 0}

  champ_podiums = c(champ_podiums, (first_place + second_place + third_place))
  champ_wins = c(champ_wins, first_place)
  num_races = c(num_races, length(subset_champ$raceId))
}

seasons = data.frame(seasons, champ_std_points, champ_podiums, champ_wins, num_races)

### RACES, WINS, PODIUMS AND RACE_POSITIONS ###

driverId = c()
races = c()
wins = c()
podiums = c()
points = c()
positions_race = c()
std_points = c()

for(i in 1:length(drivers[,1])) {
  subset_driver = results[which(results$driverId == i),]
  count_positions = subset_driver %>% count(positionOrder, sort = T)
  driverId = c(driverId, i)
  races = c(races, length(subset_driver[,1]))
```

```
  if (identical(count_positions[
    which(count_positions$positionOrder == 1), 2
  ], integer(0)) == FALSE) {
    first_place = count_positions[which(count_positions$positionOrder == 1), 2]
  } else {first_place = 0}

  if (identical(count_positions[
    which(count_positions$positionOrder == 2), 2
  ], integer(0)) == FALSE) {
    second_place = count_positions[which(count_positions$positionOrder == 2), 2]
  } else {second_place = 0}

  if (identical(count_positions[
    which(count_positions$positionOrder == 3), 2
  ], integer(0)) == FALSE) {
    third_place = count_positions[which(count_positions$positionOrder == 3), 2]
  } else {third_place = 0}

  wins = c(wins,first_place)
  podiums = c(podiums, first_place + second_place + third_place)
  points = c(points, sum(subset_driver$points))
  positions_race = c(positions_race, sum(subset_driver$positionOrder))
  std_points = c(std_points, sum(subset_driver$standarized_points))
}

data_driver = data.frame(driverId,
                         races,
                         wins,
                         podiums,
                         points,
                         positions_race,
                         std_points)
```

```
drivers = left_join(drivers, data_driver, by = c('driverId' = 'driverId'))

### PAIR WITH THE INFO WHILE THEY WERE RACING TOGETHER ###

year_function = function(x, y) {
  driver_pair = drivers[which(drivers$driverId == x | drivers$driverId == y),]

  subset_driver_x = results[which(results$driverId == x),]
  subset_driver_y = results[which(results$driverId == y),]

  years_x = subset_driver_x$year[!duplicated(subset_driver_x$year)]
  years_y = subset_driver_y$year[!duplicated(subset_driver_y$year)]

  races_together = intersect(subset_driver_x$raceId, subset_driver_y$raceId)

  subset_driver_x=subset_driver_x[which(subset_driver_x$raceId%in%races_together),]
  subset_driver_y=subset_driver_y[which(subset_driver_y$raceId%in%races_together),]

  years_together = intersect(years_x, years_y)

  if(length(races_together) == 0) {return(FALSE)} else {
    seasons_subset = seasons[which(seasons$year %in% years_together),c(8,7,6)]

    means_champ = c(mean(seasons_subset[,1]),
                    mean(seasons_subset[,2]),
                    mean(seasons_subset[,3]))
  }
  return(means_champ)
}

tabela_par = function(x, y) {
  driver_pair = drivers[which(drivers$driverId == x | drivers$driverId == y),]
```

```
subset_driver_x = results[which(results$driverId == x),]
subset_driver_y = results[which(results$driverId == y),]

years_x = subset_driver_x$year[!duplicated(subset_driver_x$year)]
years_y = subset_driver_y$year[!duplicated(subset_driver_y$year)]

races_together = intersect(subset_driver_x$raceId, subset_driver_y$raceId)

subset_driver_x=subset_driver_x[
  which(subset_driver_x$raceId%in%races_together),]

subset_driver_y=subset_driver_y[
  which(subset_driver_y$raceId%in%races_together),]

years_together = intersect(years_x, years_y)

if(length(races_together) == 0) {return(FALSE)} else {

  ### TITLES ###

  seasons_together = seasons[which(seasons$year %in% years_together),]

  count_champs_pair = seasons_together %>% count(driverRef, sort = T)
  count_champs_pair = left_join(count_champs_pair,
                                drivers %>% select(driverRef, driverId))

  if (identical(count_champs_pair$n[
      which(count_champs_pair$driverId == x)
  ], integer(0)) == FALSE) {
      driver_pair$titles[
          which(driver_pair$driverId == x)
      ] = count_champs_pair$n[which(count_champs_pair$driverId == x)]
  } else {driver_pair$titles[which(driver_pair$driverId == x)] = 0}
```

```
if (identical(count_champs_pair$n[
    which(count_champs_pair$driverId == y)
], integer(0)) == FALSE) {
    driver_pair$titles[
        which(driver_pair$driverId == y)
    ] = count_champs_pair$n[which(count_champs_pair$driverId == y)]
} else {driver_pair$titles[which(driver_pair$driverId == y)] = 0}

##### DRIVER X #####

### RACES, WINS, PODIUMS AND RACE_POSITIONS ###

count_positions = subset_driver_x %>% count(positionOrder, sort = T)

if (identical(count_positions[
    which(count_positions$positionOrder == 1), 2
], integer(0)) == FALSE) {
    first_place = count_positions[
        which(count_positions$positionOrder == 1), 2
    ]
} else {first_place = 0}

if (identical(count_positions[
    which(count_positions$positionOrder == 2), 2
], integer(0)) == FALSE) {
    second_place = count_positions[
        which(count_positions$positionOrder == 2), 2
    ]
} else {second_place = 0}

if (identical(count_positions[
    which(count_positions$positionOrder == 3), 2
```

```
], integer(0)) == FALSE) {
    third_place = count_positions[
        which(count_positions$positionOrder == 3), 2
    ]
} else {third_place = 0}

driver_pair$races[
    which(driver_pair$driverId == x)
] = length(subset_driver_x[,1])

driver_pair$wins[which(driver_pair$driverId == x)] = first_place

driver_pair$podiums[which(driver_pair$driverId == x)] = first_place +
                                                        second_place +
                                                        third_place


driver_pair$points[
    which(driver_pair$driverId == x)
] = sum(subset_driver_x$points)

driver_pair$positions_race[
    which(driver_pair$driverId == x)
] = sum(subset_driver_x$positionOrder)

driver_pair$std_points[
    which(driver_pair$driverId == x)
] = sum(subset_driver_x$standarized_points)

##### PILOTO Y #####

### RACES, WINS, PODIUMS AND RACE_POSITIONS ###

count_positions = subset_driver_y %>% count(positionOrder, sort = T)
```

```
if (identical(count_positions[
    which(count_positions$positionOrder == 1), 2
], integer(0)) == FALSE) {
    first_place = count_positions[
        which(count_positions$positionOrder == 1), 2
    ]
} else {first_place = 0}

if (identical(count_positions[
    which(count_positions$positionOrder == 2), 2
], integer(0)) == FALSE) {
    second_place = count_positions[
        which(count_positions$positionOrder == 2), 2
    ]
} else {second_place = 0}

if (identical(count_positions[
    which(count_positions$positionOrder == 3), 2
], integer(0)) == FALSE) {
    third_place = count_positions[
        which(count_positions$positionOrder == 3), 2
    ]
} else {third_place = 0}

driver_pair$races[which(driver_pair$driverId==y)]=length(subset_driver_y[,1])

driver_pair$wins[which(driver_pair$driverId==y)]=first_place

driver_pair$podiums[which(driver_pair$driverId==y)]=first_place +
                                                    second_place +
                                                    third_place
driver_pair$points[
```

```
        which(driver_pair$driverId==y)
    ]=sum(subset_driver_y$points)

    driver_pair$positions_race[
        which(driver_pair$driverId==y)
    ]=sum(subset_driver_y$positionOrder)

    driver_pair$std_points[
        which(driver_pair$driverId==y)
    ]=sum(subset_driver_y$standarized_points)

    return(driver_pair)
  }
}

exemplo_final = drivers[c(1,4,20,30,830),c(1,2,7,8,9,10,13)]

matrix_mult = function(func, drivers_final) {
  num_drivers = length(drivers_final$driverId)
  matriz_f = matrix(rep(0,num_drivers^2),num_drivers,num_drivers)
  pilotos = drivers_final$driverId

  if (func == 1) {
    cat("Função 1 escolhida\n")
    for(i in 1:num_drivers){
      cat("PILOTO ",i,"\n")
      for(j in 1:num_drivers) {
        tabela_compara = tabela_par(pilotos[i],pilotos[j])
        function_values = year_function(pilotos[i], pilotos[j])
        if(identical(tabela_compara,FALSE) == TRUE) {
          matriz_f[i,j] = 0
        } else {
          if(pilotos[i] < pilotos[j]) {
```

```
          matriz_f[i,j] = 317*tabela_compara[1,7] +
                          39.625*tabela_compara[1,9] +
                          15.85*tabela_compara[1,10] +
                          tabela_compara[1,13]
        }
        else if(pilotos[i] > pilotos[j]) {
          matriz_f[i,j] = 317*tabela_compara[2,7] +
                          39.625*tabela_compara[2,9] +
                          15.85*tabela_compara[2,10] +
                          tabela_compara[2,13]
        }
        else {
          matriz_f[i,j] = 0}
      }
    }
  }
}

if (func == 2) {
  cat("Função 2 escolhida\n")
  for(i in 1:num_drivers){
    cat("PILOTO ",i,"\n")
    for(j in 1:num_drivers) {
      tabela_compara = tabela_par(pilotos[i],pilotos[j])
      function_values = year_function(pilotos[i], pilotos[j])
      if(identical(tabela_compara,FALSE) == TRUE) {
        matriz_f[i,j] = 0
      } else {
        if(pilotos[i] < pilotos[j]) {
          matriz_f[i,j] = 317^(tabela_compara[1,7]) +
                          39.625*(tabela_compara[1,9]) +
                          15.85*tabela_compara[1,10] +
                          tabela_compara[1,13]
```

```
        }
        else if(pilotos[i] > pilotos[j]) {
          matriz_f[i,j] = 317^(tabela_compara[2,7]) +
                          39.625*(tabela_compara[2,9]) +
                          15.85*tabela_compara[2,10] +
                          tabela_compara[2,13]
        }
        else {
          matriz_f[i,j] = 0}
      }
    }
  }
}

if (func == 3) {
  cat("Função 3 escolhida\n")
  for(i in 1:num_drivers){
    cat("PILOTO ",i,"\n")
    for(j in 1:num_drivers) {
      tabela_compara = tabela_par(pilotos[i],pilotos[j])
      function_values = year_function(pilotos[i], pilotos[j])
      if(identical(tabela_compara,FALSE) == TRUE) {
        matriz_f[i,j] = 0
      } else {
        if(pilotos[i] < pilotos[j]) {
          matriz_f[i,j] = 4*317*tabela_compara[1,7] +
                          3*39.625*tabela_compara[1,9] +
                          2*15.85*tabela_compara[1,10] +
                          tabela_compara[1,13]
        }
        else if(pilotos[i] > pilotos[j]) {
          matriz_f[i,j] = 4*317*tabela_compara[2,7] +
                          3*39.625*tabela_compara[2,9] +
```

```
                        2*15.85*tabela_compara[2,10] +
                        tabela_compara[2,13]
        }
        else {
          matriz_f[i,j] = 0}
      }
    }
  }
}

if (func == 4) {
  cat("Função 4 escolhida\n")
  for(i in 1:num_drivers){
    cat("PILOTO ",i,"\n")
    for(j in 1:num_drivers) {
      tabela_compara = tabela_par(pilotos[i],pilotos[j])
      function_values = year_function(pilotos[i], pilotos[j])
      if(identical(tabela_compara,FALSE) == TRUE) {
        matriz_f[i,j] = 0
      } else {
        if(pilotos[i] < pilotos[j]) {
          matriz_f[i,j] = (317^4)*(tabela_compara[1,7]) +
                        (39.625^3)*(tabela_compara[1,9]) +
                        (15.85^2)*tabela_compara[1,10] +
                        tabela_compara[1,13]
        }
        else if(pilotos[i] > pilotos[j]) {
          matriz_f[i,j] = (317^4)*(tabela_compara[2,7]) +
                        (39.625^3)*(tabela_compara[2,9]) +
                        (15.85^2)*tabela_compara[2,10] +
                        tabela_compara[2,13]
        }
        else {
```

```
          matriz_f[i,j] = 0}
      }
    }
  }
}

if (func == 5) {
  cat("Função 5 escolhida\n")
  for(i in 1:num_drivers){
    cat("PILOTO ",i,"\n")
    for(j in 1:num_drivers) {
      tabela_compara = tabela_par(pilotos[i],pilotos[j])
      function_values = year_function(pilotos[i], pilotos[j])
      if(identical(tabela_compara,FALSE) == TRUE) {
        matriz_f[i,j] = 0
      } else {
        if(pilotos[i] < pilotos[j]) {
          matriz_f[i,j] = 317^(tabela_compara[1,7]) +
                          (39.625^3)*(tabela_compara[1,9]) +
                          (15.85*2)*tabela_compara[1,10] +
                          tabela_compara[1,13]/10
        }
        else if(pilotos[i] > pilotos[j]) {
          matriz_f[i,j] = 317^(tabela_compara[2,7]) +
                          (39.625^3)*(tabela_compara[2,9]) +
                          (15.85*2)*tabela_compara[2,10] +
                          tabela_compara[2,13]/10
        }
        else {
          matriz_f[i,j] = 0}
      }
    }
  }
```

```
}


matriz_omega = matrix(rep(0,num_driversˆ2),num_drivers,num_drivers)

for(i in 1:num_drivers) {
  for(j in 1:num_drivers) {
    if(matriz_f[i,j] == 0) {
      if(matriz_f[j,i] > 0) {
        matriz_omega[i,j] = matriz_f[j,i]
      } else {
        matriz_omega[i,j] = 0
      }
    }
    else {matriz_omega[i,j] = matriz_f[j,i]/matriz_f[i,j]}
  }
}

matriz_p = matrix(rep(0,num_driversˆ2),num_drivers,num_drivers)

for(i in 1:num_drivers) {
  for(j in 1:num_drivers) {
    matriz_p[i,j] = matriz_omega[i,j]/sum(matriz_omega[i,])
  }
}

p_opt = matriz_p
p_ant = matriz_p
count = 1
delta = 1

while (abs(delta) > 0.000001) {
  p_ant = p_opt
```

```
    p_opt = p_opt%*%matriz_p
    count = count + 1
    delta = max(p_opt - p_ant)
    cat('count = ',count,' delta = ',delta,'\n')
  }

  pi = as.numeric(p_opt[1,])
  pilot = drivers_final$driverRef
  race = drivers_final$races
  win = drivers_final$wins
  title = drivers_final$titles
  podium = drivers_final$podiums
  point = drivers_final$std_points
  rank = rank(pi)

  rank_matrix = matrix(c(pi,pilot,race,title,win,podium,point,rank),
                       ncol = 8,
                       byrow = F)

  ranking = data.frame(rank_matrix)

  return(ranking)
}

drivers_final_1 = drivers[drivers$std_points > 0,]
drivers_final_1 = drivers_final_1[1:(length(drivers_final_1$driverId)-1),]

drivers_final_2 = drivers[drivers$wins > 0 | drivers$races >= 30,]
drivers_final_2 = drivers_final_3[1:length(drivers_final_3$driverId)-1,]

drivers_final_3 = drivers[drivers$wins > 0 & drivers$races >= 30,]
drivers_final_3 = drivers_final_3[1:length(drivers_final_3$driverId)-1,]
```

```
drivers_final_4 = drivers[drivers$titles > 0,]

ranking_11 = matrix_mult(1, drivers_final_1)
ranking_12 = matrix_mult(1, drivers_final_2)
ranking_13 = matrix_mult(1, drivers_final_3)
ranking_23 = matrix_mult(2, drivers_final_3)
ranking_33 = matrix_mult(3, drivers_final_3)
ranking_43 = matrix_mult(4, drivers_final_3)
ranking_53 = matrix_mult(5, drivers_final_3)



ranking_24 = matrix_mult(2, drivers_final_4)
ranking_34 = matrix_mult(3, drivers_final_4)
ranking_44 = matrix_mult(4, drivers_final_4)
ranking_54 = matrix_mult(5, drivers_final_4)

write.csv(ranking_11, "Ranking 11.csv")
write.csv(ranking_12, "Ranking 12.csv")
write.csv(ranking_13, "Ranking 13.csv")
write.csv(ranking_23, "Ranking 23.csv")
write.csv(ranking_33, "Ranking 33.csv")
write.csv(ranking_43, "Ranking 43.csv")
write.csv(ranking_53, "Ranking 53.csv")
write.csv(ranking_24, "Ranking 24.csv")
write.csv(ranking_34, "Ranking 34.csv")
write.csv(ranking_44, "Ranking 44.csv")
write.csv(ranking_54, "Ranking 54.csv")
```