FUNDAÇÃO GETULIO VARGAS
ESCOLA DE MATEMÁTICA APLICADA

IGOR PATRÍCIO MICHELS

MODELAGEM MATEMÁTICA DO FUTEBOL BRASILEIRO: UM OLHAR NO NÍVEL INDIVIDUAL

Rio de Janeiro
2022

IGOR PATRÍCIO MICHELS

MODELAGEM MATEMÁTICA DO FUTEBOL BRASILEIRO: UM OLHAR NO NÍVEL INDIVIDUAL

Trabalho de conclusão de curso apresentada para a Escola de Matemática Aplicada (FGV/EMAp) como requisito para o grau de bacharel em Matemática Aplicada.

Área de estudo: Modelagem Matemática.

Orientador: Dr. Moacyr Alvim Horta Barbosa da Silva

Rio de Janeiro

2022

Ficha catalográfica elaborada pela BMHS/FGV

Michels, Igor
Modelagem Matemática do Futebol Brasileiro: Um olhar no nível individual/ Igor Patrício Michels. – 2022.
52f.

Trabalho de Conclusão de Curso – Escola de Matemática Aplicada.

Advisor: Dr. Moacyr Alvim Horta Barbosa da Silva.
Includes bibliography.

1. Modelagem Matemática 2. Futebol 3. Inferência Bayesiana I. Silva, Moacyr II. Escola de Matemática Aplicada III. Modelagem Matemática do Futebol Brasileiro

IGOR PATRÍCIO MICHELS

MODELAGEM MATEMÁTICA DO FUTEBOL BRASILEIRO: UM OLHAR NO NÍVEL INDIVIDUAL

Trabalho de conclusão de curso apresentada para a Escola de Matemática Aplicada (FGV/EMAp) como requisito para o grau de bacharel em Matemática Aplicada.

Área de estudo: Modelagem Matemática.

E aprovado em 06/12/2022
Pela comissão organizadora

Dr. Moacyr Alvim Horta Barbosa da Silva
FGV EMAp

Luiz Max F. de Carvalho
FGV EMAp

Walter Wagner C. Sande
FGV EMAp

Dedico essa dissertação à minha avó Ângela (*in memoriam*), que sempre me incentivou a estudar e foi fundamental para que eu estivesse aqui.

Agradecimentos

Gostaria de agradecer a todas as pessoas que contribuíram comigo durante toda essa jornada, em especial:

Meus pais, Patrício e Dilma, bem como minha irmã Natália e meu avô João, por todo incentivo aos estudos e por sempre apoiar minhas escolhas, estando sempre ao meu lado, mesmo eu estando distante de casa.

A minha namorada, Fernanda, que me apoia sempre, até mesmo nas ideias mais malucas, além de dar sua mais sincera opinião acerca das versões prévias deste trabalho.

Aos amigos que reencontrei e as novas amizades que fiz durante esse período, os quais auxiliaram tanto com suporte emocional, por estar distante de minha família, quanto com discussões sobre os tópicos visto em sala.

Ao Centro de Desenvolvimento da Matemática e Ciências da FGV, que acreditou em meu potencial e me auxiliou, tanto emocionalmente quanto financeiramente, durante esses quatro anos.

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pela concessão de uma bolsa de estudos durante meu projeto de iniciação científica, o qual culminou nesse trabalho.

Por fim gostaria de agradecer a todos os professores que, de alguma forma, contribuíram para minha formação. Em especial gostaria de agradecer a meu orientador, Moacyr Silva, por concordar em me auxiliar na condução desse trabalho, sugerindo algumas melhorias.

"A vida é como um jogo de futebol, cada lance pode definir sua trajetória."
Michael Johnathon

Resumo

O futebol é uma paixão nacional. O Brasil, apesar de não ser o berço desse esporte, é conhecido como “país do Futebol”, sendo natural que muitas conversas surjam com essa natureza. Em geral tais conversas abrangem tópicos como “qual o melhor time”, “quem irá ganhar uma determinada partida ou competição” ou então comparando dois jogadores, avaliando qual é o melhor ou o preferido para um contratação do seu clube de coração. Nesse sentido, o estudo desse esporte, a nível de modelagem, pode trazer um embasamento matemático à conversa, deixando o “achismo” de lado.

No presente trabalho analisou-se o desempenho dos jogadores das Séries A e B do campeonato brasileiro de futebol, organizado pela CBF (Confederação Brasileira de Futebol), sob o ponto de vista da Inferência Bayesiana. Para tanto, realizou-se a raspagem dos dados das súmulas de tais competições.

No contexto da modelagem três modelos foram propostos, um utilizando apenas os dados das equipes que foram a campo, bem como o resultado dessas partidas, enquanto os outros dois também se valiam da informação do mando de campo, com o intuito de modelar o diferente comportamento, dentro e fora de casa. Os principais resultados se dão pela estimação de distribuições de parâmetros que quantificam o desempenho dos jogadores, possibilitando, dessa forma, comparar e ranquear os mesmos, além de ranquear os clubes com maior influência dentro de seus domínios.

Palavras-chave: Modelagem matemática. Futebol. Inferência Bayesiana.

Abstract

Soccer is a national passion. Brazil, in spite of not being the birthplace of this sport, is known as the “country of Soccer”, and it is natural that many conversations arise with this theme. In general, such conversations cover topics such as “which team is the best”, “who will win a certain match or competition” or comparing two players, evaluating which is the best or the preferred one for signing your favorite club. In this sense, the study of this sport, at the modeling level, can bring a mathematical basis to the conversation, leaving “guessing” aside.

In this work, we analyzed the performance of players from Series A and B of the Brazilian soccer championship, organized by the CBF (Confederação Brasileira de Futebol), from the point of view of Bayesian inference. For that, the data of the summaries of such competitions was scraped.

In the context of the modeling, three models were proposed, one using only the line-up’s, as well as the result of these matches, while the other two also used information from the field command, in order to model the different behavior, inside and outside the house. The main results are given by the estimation of distributions of parameters that quantify the performance of the players, making it possible, in this way, to compare and rank them, in addition to ranking the clubs with greater influence within their domains.

Keywords: Mathematical modeling. Soccer. Bayesian inference.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

AME/MG	América Mineiro / MG
ATL/GO	Atlético Goianiense / GO
AVA/SC	Avaí / SC
BOT/RJ	Botafogo / RJ
BOT/SP	Botafogo / SP
BRA/RS	Brasil de Pelotas / RS
CAM/MG	Atlético Mineiro / MG
CBF	Confederação Brasileira de Futebol
CEA/CE	Ceará / CE
CHA/SC	Chapecoense / SC
COR/SP	Corinthians / SP
CRB/AL	CRB / AL
CRI/SC	Criciúma / SC
CSA/AL	CSA / AL
CUI/MT	Cuiabá / MT
FIDE	Fédération Internationale des Échecs (Federação Internacional de Xadrez)
FIFA	Fédération internationale de Football Association (Federação Internacional de Futebol)
FOR/CE	Fortaleza / CE
GOI/GO	Goiás / GO
GUA/SP	Guarani / SP
HMC	Hamiltonian Monte Carlo
INT/RS	Internacional / RS

JUV/RS	Juventude / RS
LON/PR	Londrina / PR
NOV/SP	Novorizontino / SP
OES/SP	Oeste / SP
OPE/PR	Operário / PR
RBB/SP	Red Bull Bragantino / SP
SAM/MA	Sampaio Côrrea / MA
SAO/SP	São Bento / SP
VIL/GO	Vila Nova / GO

Lista de símbolos

g	partida.
n_g	número de substituições de uma partida g.
k	substituições que ocorreram na partida.
$t_{g,k}$	tempo de jogo da k-ésima subpartida do jogo g.
$\mathcal{C}_{m,g,k}$	clube mandante (conjunto de jogadores) da k-ésima subpartida do jogo g.
$\mathcal{C}_{v,g,k}$	clube visitante (conjunto de jogadores) da k-ésima subpartida do jogo g.
$s_{m,g,k}$	número de gols do clube mandante na k-ésima subpartida do jogo g.
$s_{v,g,k}$	número de gols do clube visitante na k-ésima subpartida do jogo g.
$s_{m,g}$	número de gols do clube mandante na partida g.
$s_{v,g}$	número de gols do clube visitante na partida g.
p_a	vetor de proficiência de ataque dos jogadores.
p_d	vetor de proficiência de defesa dos jogadores.
$p_{i,a}$	proficiência de ataque do jogador i.
$p_{i,d}$	proficiência de defesa do jogador i.
p_{i,a_m}	proficiência de ataque do jogador i referente a jogos em casa.
p_{i,d_m}	proficiência de defesa do jogador i referente a jogos em casa.
p_{i,a_v}	proficiência de ataque do jogador i referente a jogos fora.
p_{i,d_v}	proficiência de defesa do jogador i referente a jogos fora.
$q_{i,m,g,k}$	quantidade de vezes que o "jogador i" aparece no clube mandante na k-ésima subpartida do jogo g.
$q_{i,v,g,k}$	quantidade de vezes que o "jogador i" aparece no clube visitante na k-ésima subpartida do jogo g.
$\lambda_{m,a,g,k}$	parâmetro de ataque do clube mandante na k-ésima subpartida do jogo g.

$\lambda_{v,a,g,k}$	parâmetro de ataque do clube visitante na k-ésima subpartida do jogo g.
$\lambda_{m,d,g,k}$	parâmetro de defesa do clube mandante na k-ésima subpartida do jogo g.
$\lambda_{v,d,g,k}$	parâmetro de defesa do clube visitante na k-ésima subpartida do jogo g.
$\kappa_{m,a,g,k}$	parâmetro de melhoria de ataque do clube mandante na k-ésima subpartida do jogo g.
$\kappa_{m,d,g,k}$	parâmetro de melhoria de defesa do clube mandante na k-ésima subpartida do jogo g.
G	conjunto de partidas que estão sendo modeladas.
n_p	número de jogadores considerados na modelagem.

Sumário

1 Introdução

O futebol é o esporte mais popular do mundo. Inventado na Inglaterra durante o século XIX e trazido ao Brasil por meio do paulista Charles William Miller, em 1894 (HISTÓRIA DO FUTEBOL NO BRASIL, 2022), esse esporte virou paixão nacional. Tal paixão acaba unindo pessoas de diversas classes sociais, raças, credos e regiões em virtude de uma partida ou de um título de um clube.

Além da parte social do esporte, o futebol também tem muito dinheiro envolvido. Um estudo recente da CBF em parceria com consultoria "EY" mostrou que o futebol não é mais apenas um lazer, mas também um negócio que representa 0,72% do PIB brasileiro (O ECONOMISTA, 2022). Esse montante se dá não apenas pelo valor arrecadado em dias de jogos (ingressos e toda a logística necessária para a partida), mas também com o aumento do consumo em restaurantes e bares, nos quais muitos torcedores acabam por assistir as partidas e discutir sobre o jogo com outros torcedores. Conversas como "que time é melhor?", "quem vai ser o campeão do campeonato X?" ou "o jogador A é melhor que o jogador B" surgem quase que naturalmente, muitas vezes à base do achismo.

Dentro do contexto esportivo, outro ramo que vem crescendo muito nos últimos anos é o mercado de apostas. Por ser o esporte mais popular do mundo, o mesmo corresponde a maior fatia desse mercado, que deve chegar a cerca de 180 bilhões de dólares em 2030 (GVR, 2022). Para se ter uma ideia de quão grande esse mercado já está, a Copa do Mundo de 2018, realizada na Rússia, rendeu a Fifa pouco menos de 6 bilhões de euros. Por outro lado, o volume global de apostas vinculadas a competição foi cerca de 22 vezes maior (GARCIA, 2022).

Dessa forma, o estudo de modelos matemáticos voltados ao esporte está se tornando mais relevante, tanto para levar um embasamento estatístico para conversas quanto para previsão esportiva, seja aplicada as casas de apostas ou junto ao apostador. Nesse trabalho buscamos modelar o futebol brasileiro a nível individual, isso é, buscaremos responder perguntas como "quais são os melhores jogadores que atuam no Brasil?" e "o jogador A é melhor que o jogador B?".

Esse trabalho se organiza por meio da introdução, seguida da revisão de literatura e da descrição do processo de raspagem dos dados, o qual possibilitou toda a modelagem proposta. Após o processo de raspagem seguimos para a modelagem, a qual se divide em premissas, tratamento dos dados e nas duas abordagens exploradas, a frequentista e a bayesiana.

2 Revisão de literatura

Como comentado no Capítulo 1, modelos matemáticos voltados ao estudo de esportes estão cada vez mais comuns, podendo citar exemplos de modelos voltados ao tênis (FERREIRA, 2018), hockey no gelo (BEAUDOIN; SWARTZ, 2010) e basquete (JONES, 2016), entre outros. Embora tais esportes sejam distintos, com regras completamente diferentes, alguns elementos de modelagem acabam por se repetir. (BEAUDOIN; SWARTZ, 2010), por exemplo, se vale de uma modelagem dada por modelos de Poisson para modelar os gols no hockey, classe de modelos já utilizada por Karlis (KARLIS; NTZOUFRAS, 2000) (KARLIS; NTZOUFRAS, 2003) e Lee (LEE, 2012) para o futebol.

Recentemente, Santos (SANTOS, 2019) e Maia (MAIA, 2021) estudaram modelos matemáticos voltados ao Campeonato Brasileiro. Enquanto Santos comparou os resultados de oito modelos propostos, Maia focou seu estudo na previsão dos resultados in-play, isto é, durante a partida, analisando como as probabilidades de resultado varia conforme o decorrer da partida. Novamente, tais modelos são baseados em modelos de Poisson.

Os trabalhos supracitados apresentam resultados interessantes ao realizar a modelagem do esporte ao nível de clubes, ou seja, considerando cada uma das partidas como um confronto entre duas "entidades atômicas". Cada uma dessas "entidades" possuí alguma força estimada por meio dos dados observados, o que possibilita a modelagem e a comparação entre os clubes. Entretanto, como os clubes são formados por um conjunto de jogadores, é natural pensar na modelagem a nível individual, agora considerando cada jogador como uma "entidade atômica" para o modelo.

Partindo para a modelagem individual, Wolf (WOLF; SCHMITT; SCHULLER, 2020) propõe um ranqueamento baseado no algoritmo Elo, proposto por Arpad Emrick Elo (1903 - 1992) nos anos 60 e utilizado até hoje pela FIDE. Para a modelagem, Wolf se vale dos dados de 18 competições europeias das temporadas 14/15 a 17/18. Os resultados desse artigo são interessantes, trazendo tanto o ranqueamento dos jogadores como também qual o impacto de cada jogador para seu clube.

Para o presente trabalho o objetivo é misturar algumas das ideias acima, realizando a modelagem do futebol brasileiro por meio de modelos de Poisson, mas a nível individual. Tal objetivo visa analisar o desempenho da modelagem de jogadores em relação a modelagem de clubes, além de possibilitar uma melhor interpretabilidade na predição de resultados para o futebol em relação a predição obtida com o algoritmo Elo.

3 Aquisição dos dados

Tendo em vista os modelos abordados no capítulo anterior, bem como o escopo idealizado para o trabalho, o primeiro passo se dá pela aquisição dos dados. Com isso em mente, esse capítulo visa mostrar como ocorreu tal processo.

3.1 Processo de raspagem

Como etapa inicial, realizou-se a extração dos dados dos jogos realizados em torneios da CBF (CBF, 2022) entre os anos de 2013 e 2022. O processo de raspagem dos dados foi realizado em Python, dentro de um servidor Linux, e dividido em três etapas.

A primeira etapa se refere a extração da informação propriamente dita, sendo utilizadas, principalmente, as bibliotecas `requests` e `PyPDF2`. Nessa etapa, a ideia era a de extrair as súmulas, em PDF, do site da CBF e convertê-las para `csv`. Como as súmulas já seriam acessadas e armazenadas, temporariamente, na máquina local, optou-se por guardá-las para identificação e correção de possíveis erros na próxima etapa da raspagem.

A segunda etapa, iniciada após a conversão das súmulas para `csv`, consistiu na extração das informações de cada jogo: times, resultado final, jogadores relacionados, gols e substituições. Por questões de simplicidade, as expulsões não foram coletadas. Ao final dessa etapa, as informações de cada campeonato, separadas por ano, foram convertidas para `json`, sendo codificadas pelo número do jogo, que pode ser encontrado na própria súmula. Como exemplo, a Listing 3.1 mostra como o primeiro jogo de 2022, entre Atlético Mineiro e Internacional, foi armazenado.

```
{
  "001": {
    "Mandante": "Atletico Mineiro / MG",
    "Visitante": "Internacional / RS",
    "Resultado": "2 X 0",
    "Jogadores": [
      [
        "22Everson Everson Felipe Marqu ... T(g)P188398",
        "Atletico Mineiro / MG"
      ],
      [
        "2Guga Claudio Rodrigues Gomes TP317445",
        "Atletico Mineiro / MG"
      ],
```

```
      ],
      ...,
      [
        "1Daniel Daniel de Sousa Brito T(g)P331033",
        "Internacional / RS"
      ],
      [
        "3Kaique Kaique Rocha Lima TP442170",
        "Internacional / RS"
      ],
      ...
    ],
    "Gols": [
      "09:00 1T 7 NR Givanildo Vieira de Sousa Atletico
         Mineiro/MG",
      "+2 2T 7 NR Givanildo Vieira de Sousa Atletico Mineiro/
         MG"
    ],
    "Substituicoes": [
      "45:00 INT Internacional/RS 25 - Gabriel Ivan Mercado 9
          - Wesley Moraes Ferreira da Si...",
      "12:00 2TAtletico Mineiro/MG 15 - Federico Matias
         Zaracho 18 - Eduardo Colcenti Antunes",
      "19:00 2TAtletico Mineiro/MG 8 - Jair Rodrigues Junior
         5 - Otavio Henrique Passos Santos",
      "19:00 2TAtletico Mineiro/MG 17 - Jefferson David
         Savarino Qui... 19 - Ademir da Silva Santos Junior",
      "20:00 2TInternacional/RS 35 - Alexandre Zurawski 27 -
         Mauricio Magalhaes Prado",
      "33:00 2TAtletico Mineiro/MG 10 - Eduardo Jesus Vargas
         Rojas 11 - Marcos da Silva Franca",
      "34:00 2TInternacional/RS 21 - Gabriel Boschilia 5 -
         Igor Matheus Liziero Pereira",
      "43:00 2TInternacional/RS 47 - Caio Vidal Rocha 22 -
         Bruno Mendez Cittadini"
    ]
  }
}
```

Listing 3.1 – Exemplo de JSON obtido após a segunda etapa.

Tendo concluído a segunda etapa, partiu-se para a montagem das escalações de cada um dos jogos. Para tanto, os jogadores foram identificados por seu código CBF. Nessa etapa, além da escalação inicial (os vinte e dois atletas que iniciaram a partida), foram montadas escalações secundárias, de modo que cada uma das escalações representasse um momento da partida, isso é, o período entre uma substituição e outra. Assim, a primeira escalação se deu pelos times que entraram em campo, enquanto as demais foram elaboradas seguindo as próprias alterações da partida, removendo o jogador que foi substituído e incluindo o atleta que entrou em campo. Para substituições que ocorrem no mesmo minuto, como a entrada de Jair e Jefferson para a saída de Otávio e Ademir, pelo Atlético Mineiro, observada na Listing 3.1, foi considerada apenas uma nova escalação.

Consequentemente, com essa quebra do jogo em função das escalações, o tempo de jogo e o placar também foram adaptados para coincidir com o intervalo de jogo que a escalação atual se refere. Com o objetivo de ilustrar essa ideia, a Listing 3.2 traz o início do arquivo das escalações, mostrando as escalações elaboradas para o jogo citado previamente. Novamente, o primeiro indexador do arquivo é o código do jogo, dado pela CBF, e o segundo pode ser interpretado como o número de paradas para substituição de jogadores que ocorreram até o início daquela escalação.


```
{
  "001": {
    "0": {
      "Mandante": [
        "188398", "317445", "699115", "320930", "163056", "375634", "339258", "295792", "438212", "339266", "370997"
      ],
      "Visitante": [
        "331033", "442170", "344934", "177042", "306270", "387944", "756803", "750897", "648335", "320839", "542337"
      ],
      "Tempo": 45,
      "Placar": [
        1,
        0
      ]
    },
    "1": {
      "Mandante": [
        "188398", "317445", "699115", "320930", "163056", "37
```

```
            5634", "339258", "295792", "438212", "339266", "37
            0997"
19        ],
20        "Visitante": [
21          "331033", "442170", "344934", "177042", "306270", "75
              6803", "750897", "648335", "320839", "542337", "72
              3593"
22        ],
23        "Tempo": 12,
24        "Placar": [
25          0,
26          0
27        ]
28      },
29      "2": {
30        "Mandante": [
31          "188398", "317445", "699115", "320930", "163056", "37
              5634", "339258", "438212", "339266", "370997", "70
              3937"
32        ],
33        "Visitante": [
34          "331033", "442170", "344934", "177042", "306270", "75
              6803", "750897", "648335", "320839", "542337", "72
              3593"
35        ],
36        "Tempo": 7,
37        "Placar": [
38          0,
39          0
40        ],
41        ...
42      }
43    }
44  }
```

Listing 3.2 – Exemplo do JSON final.

Como comentado anteriormente, os tempos de jogo e gols também foram adaptados para a quebra das escalações. Dessa forma, note que, pela Listing 3.1, podemos ver que o Internacional sofreu um gol no primeiro tempo e que realizou uma substituição no

intervalo, sendo a primeira substituição do jogo. Esses dois eventos se refletem na Listing 3.2 com a primeira escalação (indexada pelo 0), tendo o placar de 1x0 e tempo total de 45 minutos. Como aos 12 minutos do segundo tempo o Atlético fez uma alteração, a escalação indexada pelo 1 possui 12 minutos de duração e placar igual a 0x0, pois não houveram gols nesse período. As demais escalações foram definidas de modo análogo, com o tempo de jogo limitado a 90 minutos e, todos os eventos que ocorreram durante os acréscimos, levados para o último minuto do respectivo tempo de jogo.

3.2 Problemas durante a raspagem

Durante a raspagem algumas súmulas fugiram do padrão, principalmente as que se referiam do período de 2013 a 2016. Algumas em virtude de W.O., outras com preenchimento incorreto. Os erros mais básicos, como o W.O. e a falta da unidade federativa no nome de alguns times, foram tratados de forma semi-manual e colocados em um arquivo `json` de exceções. Tendo resolvido tais problemas, a transição da primeira para a segunda etapa foi concluída, restando, portanto, apenas a outra transição.

Mais uma vez, durante o caminho da segunda para a terceira etapa foram evidenciados alguns erros, principalmente no preenchimento das súmulas, com informações que não faziam sentido. Em virtude disso, tais erros foram tratados, predominantemente, de forma manual. A opção pelo tratamento manual, apesar de trabalhosa, se deu pelo fato de que existiam erros que necessitavam de verificação externa para validar (ou até mesmo coletar) a informação correta.

Um exemplo de como tais erros ocorreram pode ser visto nas Figuras 1 e 2. Note que o jogador Centurión, do São Paulo, foi, de acordo com a súmula, relacionado como titular. Entretanto, aos 24 minutos do segundo tempo ele é chamado para entrar em campo substituindo o camisa 25, Hudson.

Outro jogador, na mesma partida, também teve sua participação documentada de forma incorreta. O camisa 11, Alexandre Pato, foi relacionado como reserva. Entretanto, aos 29 minutos do segundo tempo, o mesmo deixa o campo para a entrada de Thiago Henrique, camisa 23, sendo que, pela súmula, o camisa 11 nem teria entrado em campo previamente.

Esses erros, envolvendo a escalação inicial do clube, foram recorrentes, com mais jogos apresentando situações similares.

Relação de Jogadores

Palmeiras / SP

Nº	Apelido	Nome Completo	T/R	P/A	CBF
1	Fernando	Fernando Buttenbende ...	T	P	137377
3	Victor Ramos	Victor Ramos Ferreira	T	P	177152
5	Marcos	Marcos Arouca da Silva	T	P	161941
7	Dudu	Eduardo Pereira Rodr ...	T	P	291738
17	Leandro	Leandro Marcos Pereira	T	P	306149
18	Gabriel	Gabriel Girotto Franco	T	P	320839
19	Rafael Mar ...	Rafael Marques Mariano	T	P	158615
27	Robinho	Robson Michael Signorini	T	P	172189
31	Vitor Hugo	Vitor Hugo Franchesc ...	T	P	300140
32	Lucas	Lucas Rios Marques	T	P	178827
66	Egidio	Egidio de Araujo Per ...	T	P	157280
2	Omar Ferna ...	Omar Fernando Tobio	R	P	507274
4	Nathan	Nathan Raphael Pelae ...	R	P	317706
6	Joao Paulo	Joao Paulo Gomes da ...	R	P	155855
8	Cleiton Xa ...	Cleiton Ribeiro Xavier	R	P	154144
9	Cristaldo	Jonatan Ezequiel Cri ...	R	P	511616
11	Ze Roberto	Jose Roberto da Silv ...	R	P	120815
15	Amaral	William Jose de Souza	R	P	169939
22	Joao	Joao Pedro Maturano ...	R	P	376586
25	Aranha	Mario Lucio Duarte Costa	R	P	142720
28	Andrei	Andrei Girotto	R	P	330966

São Paulo / SP

Nº	Apelido	Nome Completo	T/R	P/A	CBF
1	Rogerio	Rogerio Ceni	T	P	107639
2	Toloi	Rafael Toloi	T	P	184801
5	Souza	Josef de Souza Dias	T	P	178096
6	Carlinhos	Carlos Andrade Souza	T	P	173760
7	Michel Bas ...	Michel Fernandes Bastos	T	P	155330
9	Luis Fabiano	Luis Fabiano Clemente	T	P	135687
10	Paulo Henr ...	Paulo Henrique Chaga...	T	P	184364
20	Centurion	Ricardo Adrian Centurion	T	P	522233
22	Bruno	Bruno Vieira do Nasc...	T	P	172996
25	Hudson	Hudson Rodrigues dos...	T	P	176729
26	Matheus	Matheus Doria Macedo	T	P	317379
4	Matheus Reis	Matheus Reis de Lima	R	P	337413
11	Alexandre ...	Alexandre Rodrigues ...	R	P	177588
17	Cafu	Jonathan Renato Barbosa	R	P	302451
19	Wesley	Wesley Lopes Beltrame	R	P	170540
21	Edson	Edson Jose da Silva	R	P	166943
23	Thiago Men ...	Thiago Henrique Mend...	R	P	353262
27	Auro	Auro Alvaro da Cruz ...	R	P	321561
28	Renan	Renan Ribeiro	R	P	176915

Figura 1 – Relação de jogadores da partida entre Palmeiras e São Paulo pela Série A de 2015.

Substituições

Tempo	1T/2T	Equipe	Entrou	Saiu
14:00	2T	Palmeiras/SP	9 - Jonatan Ezequiel Cristald ...	17 - Leandro Marcos Pereira
24:00	2T	São Paulo/SP	20 - Ricardo Adrian Centurion	25 - Hudson Rodrigues dos Sant ...
29:00	2T	São Paulo/SP	23 - Thiago Henrique Mendes Ri ...	11 - Alexandre Rodrigues da Si ...
31:00	2T	Palmeiras/SP	8 - Cleiton Ribeiro Xavier	27 - Robson Michael Signorini
36:00	2T	Palmeiras/SP	33 - Gabriel Fernando de Jesus ...	7 - Eduardo Pereira Rodrigues ...

Figura 2 – Substituições da partida entre Palmeiras e São Paulo pela Série A de 2015.

Além dos erros de preenchimento do time titular, outro erro comum foi no preenchimento das substituições, como pode ser visto na Figura 3. Nela o mesmo jogador, João Rodrigues, entrou e deixou o campo no mesmo minuto. O problema, nesse caso, se deu pelo fato dele primeiro sair de campo para, depois, entrar.

Substituições

Tempo	1T/2T	Equipe	Entrou	Saiu
12:00	2T	Paysandu	17 - Tallys Machado de Oliveir ...	19 - Joao Rodrigues Filho Neto ...
12:00	2T	Paysandu	19 - Joao Rodrigues Filho Neto ...	8 - Diego Henrique Barboza do ...
22:00	2T	América	15 - Dorival das Neves Ferraz ...	8 - Andrei Girotto
26:00	2T	América	16 - Leandro Ferreira Carvalho ...	11 - Kleber do Amaral Ribeiro
28:00	2T	América	19 - Tiago Alves Sales	7 - Willians dos Santos Santa ...
31:00	2T	Paysandu	18 - Alex Jose Arcanjo dos San ...	5 - Esdras Silva E Silva

Figura 3 – Relação de jogadores da partida entre Paysandu e América pela Série B de 2013.

Tais erros foram corrigidos manualmente, e, por simplicidade, foram alterados diretamente nos arquivos `csv` gerados na primeira etapa. Dessa forma, ao executar o processo de raspagem, a segunda etapa da extração corrige os jogos com erros nos arquivos `json` e, com isso, a passagem da segunda etapa para a terceira está corrigida.

3.3 Automação do processo

Durante o desenvolvimento do trabalho muitos campeonatos de 2022 ainda estavam em andamento, logo a extração de dados deveria, idealmente, ser constante. Aliando isso com o fato de que, no momento em que o código da raspagem foi finalizado, nenhum jogo dos últimos anos apresentou problema ao extrair as informações, poderíamos apenas executar o código da raspagem dos dados com certa frequência. Pensando em evitar que essa rotina fosse realizada de forma manual foi criado, no GitHub Actions, uma rotina para que todo dia a raspagem fosse executada. Essa rotina englobou todo o processo de raspagem, desde a coleta das súmulas até os tratamentos das etapas finais, onde o arquivo `json` é elaborado.

Nesse ponto, com o intuito de poupar custos computacionais, vale ressaltar que os arquivos finais só eram regerados no caso em que havia a inclusão de um novo jogo na base ou a alteração de algum jogo já existente. Tal regra foi criada pois se não houvesse nenhuma alteração nos dados da base as informações a serem coletadas deveriam ser, necessariamente, as mesmas.

Como o processo de raspagem requer conexão com a internet, foi elaborado um gatilho para parar a execução pois, a depender da rede, a conexão poderia demorar. Nesse sentido, foram utilizadas as bibliotecas `multiprocessing` e `time` para limitar o tempo de raspagem de cada competição, encerrando o processo caso o tempo total de extração, de uma temporada, ultrapasse dois minutos. A escolha de tal tempo foi dada em virtude de que, como a rotina é diária, o maior número de jogos que seriam coletados em um dia, para uma competição, ocorreria no caso de rodada cheia, com 10 jogos.

Entretanto, caso o tempo do processo tenha ultrapassado esse limite, mas tenham sido adicionados novos jogos, o processo de raspagem é repetido, evitando que, no caso de uma raspagem em massa (do campeonato inteiro, por exemplo), algum jogo não seja coletado em virtude da expiração do tempo.

Por fim, para acompanhamento desse processo, foi elaborado um arquivo de log, onde, como pode-se ver na Figura 4, traz as informações gerais sobre o número de jogos adicionados, além de informar caso algum jogo tenha gerado algum erro.

```
2022-09-08 07:47:18 [Scraping]   - Iniciando extração das súmulas.
2022-09-08 07:48:32 [Scraping]   - Finalizada extração das súmulas.
                                   [INFO] 3 súmulas foram adicionadas.
2022-09-08 07:48:32 [Extração]   - Iniciada extração das informações.
2022-09-08 07:48:33 [Extração]   - Finalizada extração das informações.
                                   [INFO] Todos os jogos foram extraídos com sucesso!
2022-09-08 07:48:33 [Escalações] - Iniciada geração das escalações.
2022-09-08 07:48:33 [Escalações] - Finalizada geração das escalações.

--------------------------------------------------------------------------------

2022-09-09 07:41:31 [Scraping]   - Iniciando extração das súmulas.
2022-09-09 07:42:21 [Scraping]   - Finalizada extração das súmulas.
                                   [INFO] 2 súmulas foram adicionadas.
2022-09-09 07:42:21 [Extração]   - Iniciada extração das informações.
2022-09-09 07:42:22 [Extração]   - Finalizada extração das informações.
                                   [INFO] Todos os jogos foram extraídos com sucesso!
2022-09-09 07:42:22 [Escalações] - Iniciada geração das escalações.
2022-09-09 07:42:22 [Escalações] - Finalizada geração das escalações.

--------------------------------------------------------------------------------
```

Figura 4 – Log do processo de raspagem.

4 Modelagem proposta

4.1 Premissas utilizadas

Assim como utilizado por outros autores ((BEAUDOIN; SWARTZ, 2010), (KARLIS; NTZOUFRAS, 2000), (KARLIS; NTZOUFRAS, 2003), (LEE, 2012), entre outros), estaremos considerando que os gols em uma partida de futebol seguem uma distribuição de Poisson. Para dar mais credibilidade a premissa, podemos ver, na Figura 5, a distribuição de gols por partida, separada em mandante e visitante, para todo período extraído. Para comparar com a distribuição de Poisson, plotou-se juntamente do histograma de gols, a função de massa de probabilidade da Poisson de máxima verossimilhança dos dados observados, isso é, onde $\lambda_m = \frac{1}{|G|}\sum_{g\in G} s_{m,g}$ e $\lambda_v = \frac{1}{|G|}\sum_{g\in G} s_{v,g}$.

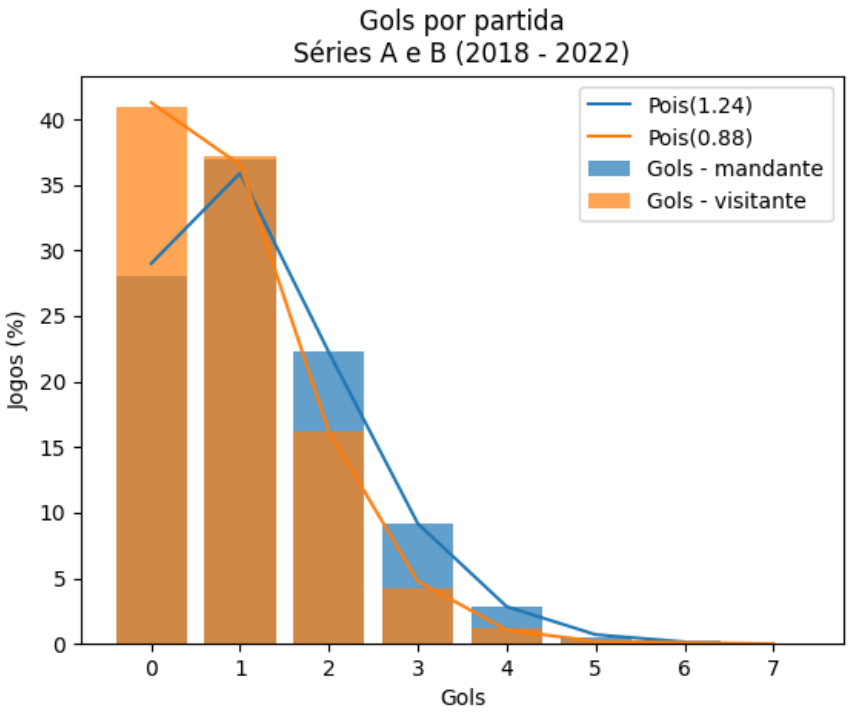


Figura 5 – Histograma de gols por partida.

Conforme podemos ver na Figura 5, as duas distribuições (gols marcados na partida e Poisson) são bem similares, embora isso não assegure que podemos afirmar que os dados vêm de uma distribuição de Poisson. Para testar essa hipótese podemos realizar um Teste χ^2 de Aderência (ou teste de Bondade do Ajuste) (NIST/SEMATECH, 2012). Esse teste pode ser utilizado com o objetivo de testar a hipótese H_0, de que o conjunto de dados observados segue uma determinada distribuição, contra a hipótese H_a, de que o conjunto de dados observados não é oriundo dessa distribuição. A estatística a ser testada é dada

por

$$\Lambda = \sum_{i \in \mathbb{Z}_+} \mathbb{1}_{O_i \neq 0} \frac{(O_i - E_i)^2}{E_i}, \tag{4.1}$$

onde $\mathbb{1}_{O_i \neq 0}$ é a função indicadora, valendo 0 se $O_i = 0$ e 1 caso contrário, O_i é o número de jogos em que o time analisado marcou i gols e E_i é o número esperado, segundo a distribuição de H_0 do número de jogos em que o time analisado marcaria i gols.

Tal estatística segue uma distribuição χ^2_{k-2} sendo,

$$k = \sum_{i \in \mathbb{Z}_+} \mathbb{1}_{O_i \neq 0}.$$

Por fim, H_0 é rejeitada caso $\Lambda > \chi^2_{1-\alpha,k-2}$, onde α é nível do teste e $\chi^2_{1-\alpha,k-2}$ é o valor crítico para o teste.

Todas as contas do teste foram realizadas em Python. Utilizando $\alpha = 0.05$, chegamos nos resultados da Tabela 1. Ambos os testes falharam em rejeitar a hipótese nula, de que os dados observados seguem a distribuição de Poisson vista na Figura 5.

Distribuição	Λ	Valor Crítico
Mandante	7.0253	11.0705
Visitante	10.4619	11.0705

Tabela 1 – Resultados do teste de Bondade do Ajuste.

Outra premissa muito importante, imposta por questões de simplicidade, é a de que os jogadores são independentes. Dessa forma, as escalações dos dois clubes são independentes e, dentro de um mesmo clube, não existem jogadores com mais ou menos entrosamento entre si. Isso é, não estamos considerando as possibilidade de haver um racha no elenco ou um grande entrosamento entre dois jogadores. Também por questões de simplicidade, não estaremos considerando o cansaço dos jogadores, assim os jogadores mantém seu rendimento igual por toda a partida. Abordaremos mais a fundo essas premissas na Subseção 4.2.2.

Por fim, como comentado brevemente no capítulo anterior, estaremos supondo que todas as partidas são independentes entre si e possuem, exatamente, 90 minutos. Além disso, todos os eventos ocorridos durante os acréscimos são considerados como se tivessem ocorrido no minuto imediatamente anterior (45 minutos do primeiro ou do segundo tempo). Tal suposição se dá pelo próprio processo de raspagem, no qual não foram considerados os acréscimos, jogando todos os eventos dos mesmos para o último minuto da etapa atual do jogo.

4.2 Tratamento e Seleção dos dados

Tendo em mãos os dados para serem utilizados e as premissas, o próximo passo é, naturalmente, a seleção dos dados que irão entrar no modelo.

4.2.1 Tratamento dos dados

Como estamos modelando as partidas em função dos jogadores que estão em campo, sempre que ocorre uma substituição todo o contexto muda, uma vez que, pelo menos, um clube mudou. Desse modo, para modelar esse fenômeno podemos redefinir as partidas, quebrando-as em subpartidas correspondentes a cada substituição. Dessa forma, uma partida g com n_g substituições será quebrada em $n_g + 1$ partidas, de modo que cada intervalo de quebra represente os eventos que ocorreram naquele intervalo. Tal quebra já foi relatada na Seção 3, além de ter sido implementada no processo de raspagem. Naturalmente, devemos ter

$$\sum_{k=0}^{n_g} t_{g,k} = 90,$$

$$\sum_{k=0}^{n_g} s_{m,g,k} = s_{m,g} \text{ e}$$

$$\sum_{k=0}^{n_g} s_{v,g,k} = s_{v,g}.$$

Exemplo 1

Em uma partida hipotética g na qual ocorreram 3 substituições, uma aos $29'$, outra aos $57'$ e outra aos $62'$, com gols aos $11'$ para o mandante e aos $87'$ para o visitante terá as seguintes características:

Nome	Valor
n_g	3
$t_{g,0}$	29
$s_{g,m,0}$	1
$s_{g,v,0}$	0
$t_{g,1}$	28
$s_{g,m,1}$	0
$s_{g,v,1}$	0
$t_{g,2}$	5
$s_{g,m,2}$	0
$s_{g,v,2}$	0
$t_{g,3}$	28
$s_{g,m,3}$	0
$s_{g,v,3}$	1
$s_{g,m}$	1
$s_{g,v}$	1

Com tais substituições, as forças gerais dos dois clubes acabam se alterando conforme o jogo acontece e os jogadores são substituídos, o que faz com que surjam os conjuntos de jogadores a cada período de jogo[1] e, consequentemente, as forças de ataque e defesa de cada clube para cada período de tempo.

4.2.2 Seleção dos dados

O objetivo do trabalho é a modelagem do futebol brasileiro atual. Assim, para a modelagem, iremos nos valer dos dados mais recentes possíveis e sempre das divisões superiores. Nesse sentido, a preferência sempre será dada a jogos das Séries A e B, com os anos mais recentes.

Pelo fato dos clubes das Séries C e D não enfrentarem todos os outros clubes de sua divisão, os dados dessas duas divisões não serão utilizados, uma vez que isso pode dificultar a modelagem de alguns jogadores, além de aumentar a complexidade do modelo. Já os dados da Copa do Brasil serão deixados de lado visando evitar os clubes das Séries C e D que jogam tal competição e, principalmente, em razão da complexidade do modelo. Assim, os dados que serão utilizados tratam das equipes das Série A e, talvez, B, restando definir o período que será considerado.

Para melhor escolha dos dados, a Tabela 2 apresenta a quantidade geral de jogadores e subpartidas que forem coletados. Como explicado anteriormente, queremos focar nos dados mais recentes de, no máximo, duas divisões do campeonato brasileiro (Séries A e B). Dessa forma, as tabelas foram organizadas de modo que os valores se acumulam à medida que avancemos para baixo e para a direita. Ou seja, o valor 3429 da primeira tabela, na linha correspondente ao ano de 2017 e da coluna correspondente a Série B significa que 3429 jogadores entraram em campo em jogos das Séries A e B entre os anos de 2017 e 2022.

Ano	Série A	Série B
2022	747	1501
2021	1048	1978
2020	1289	2415
2019	1531	2801
2018	1765	3154
2017	1982	3496
2016	2213	3901
2015	2415	4287
2014	2648	4709
2013	2851	5090

Ano	Série A	Série B
2022	2506	5059
2021	4978	10048
2020	7500	15112
2019	9864	19901
2018	12266	24725
2017	14637	29487
2016	16982	34224
2015	19377	39005
2014	21726	43711
2013	24064	48452

Tabela 2 – Dados totais. À esquerda referente ao número de jogadores, à direita referente ao total de subpartidas.

[1] Cada período sendo dado no intervalo de tempo entre substituições ou início/término da partida.

Com os dados da Tabela 2 podemos voltar a discussão iniciada na Seção 4.1, sobre a independência dos jogadores. Se considerarmos que os jogadores possuem influência um sobre o outro, devemos ter, na pior das hipóteses, para cada par de jogadores um parâmetro que represente essa influência. Tal parâmetro deve agir de modo que aumente (ou diminua) as chances de um clube marcar gol. Um exemplo simples de como isso funcionaria seria pensar em um atacante e um goleiro que, em algum momento do passado, foram companheiros de time, treinando e jogando juntos. Naturalmente o goleiro irá conhecer o atacante e vice-versa, o que possibilitaria a adoção de alguma estratégia (individual ou coletiva), para neutralizar ou explorar o adversário.

Nesse sentido, a dependência dos jogadores parece ser favorável a modelagem, ao menos na tarefa de tentar prever resultados. Porém, pela Tabela 3, que mostra a quantidade de pares de jogadores que estavam em campo simultaneamente, podemos ver que a inclusão de algum parâmetro que correlacione os jogadores deixa o modelo demasiadamente complexo. Além disso, o modelo fica inviabilizado, uma vez que o número de pares de jogadores supera, em muito, o número de subpartidas, dadas na Tabela 2. Logo, além de simplificar a modelagem, a premissa da independência dos jogadores possibilita a realização da mesma.

Ano	Série A	Série B
2022	82841	169001
2021	144390	225203
2020	201045	284071
2019	249634	318987
2018	298906	368844
2017	345388	413307
2016	393498	463228
2015	439449	509984
2014	487534	558736
2013	532641	602980

Tabela 3 – Número de pares de jogadores que estiveram em campo simultaneamente.

Por fim, por meio da observação da Tabela 2 e por testes em relação aos limites computacionais, optou-se pela utilização dos dados das Séries A e B, a partir de 2018.

4.3 Abordagem Frequentista

Como visto ao final da Seção 4.2.1, as forças de ataque e defesa dos clubes, em cada período de um jogo, irão variar, sendo dadas, em cada subpartida, pelas equações em 4.2.

$$\begin{aligned}
\lambda_{m,a,g,k} &= \sum_{i \in \mathcal{C}_{m,g,k}} p_{i,a}, \\
\lambda_{v,a,g,k} &= \sum_{i \in \mathcal{C}_{v,g,k}} p_{i,a}, \\
\lambda_{m,d,g,k} &= \sum_{i \in \mathcal{C}_{m,g,k}} p_{i,d} \text{ e} \\
\lambda_{v,d,g,k} &= \sum_{i \in \mathcal{C}_{v,g,k}} p_{i,d},
\end{aligned} \tag{4.2}$$

onde $p_{i,a}$ e $p_{i,d}$ representam, respectivamente, as proficiências de ataque e de defesa do jogador i. Notemos que aqui não faz sentido que tais parâmetros sejam negativos, uma vez que, nesse caso, a escalação de um jogador pioraria o desempenho do time. Por fim, a média da distribuição de Poisson será dado pelo produto do tempo da subpartida e a razão entre a força de ataque de um clube e a força de defesa do outro, o que representa a taxa de espera entre gols. Em termos matemáticos, podemos expressar o modelo como em 4.3.

$$\begin{aligned}
s_{m,g,k} &\sim \text{Poisson}\left(\frac{\lambda_{m,a,g,k}}{\lambda_{v,d,g,k}} \cdot t_{g,k}\right) \\
s_{v,g,k} &\sim \text{Poisson}\left(\frac{\lambda_{v,a,g,k}}{\lambda_{m,d,g,k}} \cdot t_{g,k}\right).
\end{aligned} \tag{4.3}$$

Notemos que o modelo dado pelas equações em 4.3 possui um problema de identificabilidade. Dado o conjunto de parâmetros $\mathcal{P}$, o conjunto $\mathcal{P}' = c \cdot \mathcal{P}$, definido por meio da multiplicação de cada parâmetro por c, o modelo que se vale do conjunto $\mathcal{P}'$ é idêntico ao modelo que usa o conjunto $\mathcal{P}$. Essa equivalência ocorre pelo fato do fator c ser simplificado no cálculo do parâmetro de cada uma das distribuições. Para tratar tal problema, fixou-se o primeiro parâmetro do modelo como sendo igual a 1, isso é, o parâmetro de ataque do primeiro jogador do modelo será fixo.

Agora, com o modelo definido e o problema de identificabilidade tratado, podemos iniciar o processo de modelagem. Como no contexto frequentista o objetivo é o de estimar os parâmetros que, dadas as partidas observadas, maximizam a verossimilhança do processo como um todo, devemos, primeiramente, encontrar a função de verossimilhança.

Por se tratar de várias subpartidas independentes, a verossimilhança se dá por um produtório de vários termos. Entretanto, todos esses termos podem ser interpretados como uma probabilidade, ou seja, estão em $[0, 1]$ ou, quase certamente, em $(0, 1)$. Com isso, computacionalmente, poderíamos ter problemas com ponto flutuante conforme o produtório é realizado, de modo que podemos obter um valor muito baixo, sendo impossível de representar o mesmo computacionalmente. O tratamento clássico para tal problema

é a aplicação do logaritmo que, por ser uma função monótona, não afeta o processo de otimização, isso é, como temos que $x < y \implies \log x < \log y$, se x é o valor ótimo da verossimilhança, $\log x$ será o valor ótimo da função de log-verossimilhança.

Além de tratar o problema de ponto flutuante, a aplicação do log faz com que os produtórios sejam reescritos como somatórios, o que também auxilia no tratamento do problema a nível computacional. Diante do exposto, podemos escrever a função de log-verossimilhança dos dados, conforme Equação 4.4.

$$\begin{aligned}\mathcal{L}(p_a, p_d|G) = \sum_{g}\sum_{k=0}^{n_g}\Bigg(&-\frac{\lambda_{m,a,g,k}}{\lambda_{v,d,g,k}}\cdot t_{g,k} + s_{m,g,k}\cdot\log\left(\frac{\lambda_{m,a,g,k}}{\lambda_{v,d,g,k}}\cdot t_{g,k}\right) - \log\left(s_{m,g,k}!\right)\\ &-\frac{\lambda_{v,a,g,k}}{\lambda_{m,d,g,k}}\cdot t_{g,k} + s_{v,g,k}\cdot\log\left(\frac{\lambda_{v,a,g,k}}{\lambda_{m,d,g,k}}\cdot t_{g,k}\right) - \log\left(s_{v,g,k}!\right)\Bigg).\end{aligned} \tag{4.4}$$

A função em 4.4 está bem definida, uma vez que $\lambda_{m,a,g,k}, \lambda_{m,d,g,k}, \lambda_{v,a,g,k}, \lambda_{v,d,g,k} > 0$ e $t_{g,k}, s_{m,g,k}, s_{v,g,k} > 0$. Entretanto seu domínio está em $\mathbb{R}_+^{2n_p}$, o qual possui uma dimensão elevada. Para atenuar essa complicação do domínio, sem a perda de muita informação, podemos pensar na geração de "*clusters*" de jogadores. A ideia desses "*clusters*" é de agrupar jogadores com poucas partidas disputadas como se fossem o mesmo jogador. Assim, o espaço de parâmetros diminui e, consequentemente, a complexidade do problema de otimização também, uma vez que o número de parâmetros que devem ser estimados será menor.

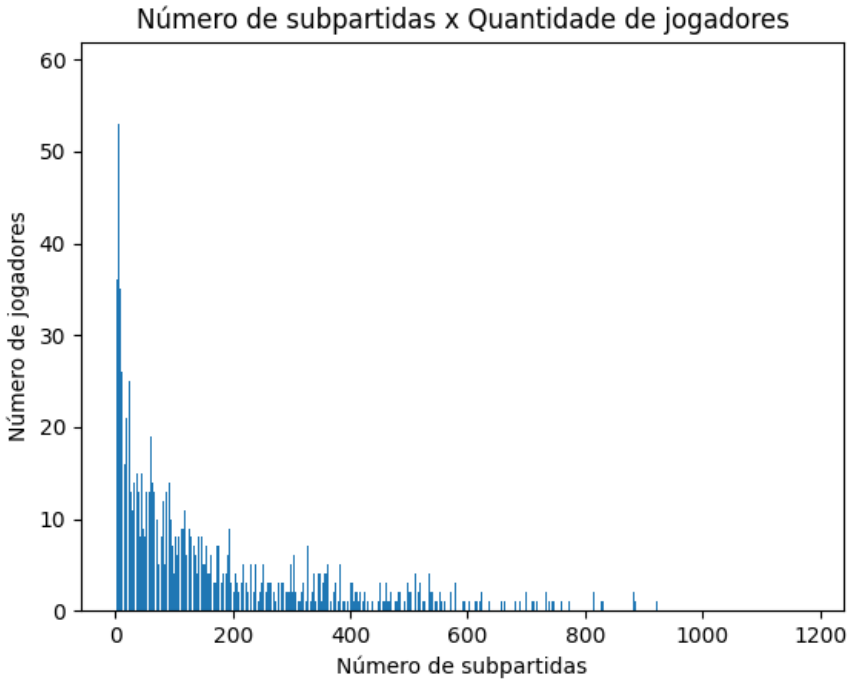


Figura 6 – Número de jogadores conforme o número de subpartidas em que ele atuou.

Para ilustrar a ideia da clusterização podemos ver a Figura 6, onde percebe-se que muitos atletas atuaram em poucas subpartidas. Nesse cenário, a função de log-

verossimilhança permanece tendo a mesma forma que vimos previamente, a única alteração se dá na definição das forças de ataque e defesa de cada clube, as quais passam a se expressar como 4.5.

$$\begin{aligned}
\lambda_{m,a,g,k} &= \sum_i q_{i,m,g,k} \cdot p_{i,a}, \\
\lambda_{v,a,g,k} &= \sum_i q_{i,v,g,k} \cdot p_{i,a}, \\
\lambda_{m,d,g,k} &= \sum_i q_{i,m,g,k} \cdot p_{i,d} \text{ e} \\
\lambda_{v,d,g,k} &= \sum_i q_{i,v,g,k} \cdot p_{i,d}.
\end{aligned} \tag{4.5}$$

Essa clusterização, apesar de simples, reduz de modo considerável o número de parâmetros. Se utilizarmos como quantidade mínima 20 subpartidas, temos que o número de jogadores a serem utilizados para a modelagem é de 2547, contra os 3154 originais, o que corresponde a uma diminuição de aproximadamente 20% no número de parâmetros a serem estimados. Nesse cenário, os jogadores com menos de 20 subpartidas são agrupados, sendo substituídos por 20 novos jogadores, cada um desses correspondendo ao conjunto de jogadores que jogou cada quantidade de subpartidas entre 1 e 20.

Outra agregação possível, agora em relação ao tempo de jogo, seria o agrupamento de jogadores com uma certa quantidade de minutos de jogo, a qual não será explorada no trabalho, mas que segue raciocínio análogo ao descrito acima.

Por fim, observe que se $q_{i,m,g,k}, q_{i,v,g,k} \in \{0, 1\}$ para todos os valores de g e k então estamos no caso anterior, onde a clusterização não foi aplicada, logo, essa é uma formulação mais geral que a inicial. Sendo assim, iremos utilizar esse modelo daqui em diante.

Apesar da clusterização juntar alguns atletas e diminuir o número de parâmetros a serem estimados, tal número permanece elevado. Nesse sentido, é interessante conhecer um pouco mais a função que desejamos otimizar. Para tanto, podemos calcular o gradiente da função, o qual pode auxiliar nos algoritmos de otimização. As derivadas parciais de $\mathcal{L}$ podem ser expressas por

$$\frac{\partial \mathcal{L}}{\partial p_{i,a}} = \sum_g \sum_{k=0}^{n_g} \left(-t_{g,k} \cdot \frac{q_{i,m,g,k}}{\lambda_{v,d,g,k}} + s_{m,g,k} \cdot \frac{q_{i,m,g,k}}{\lambda_{m,a,g,k}} - t_{g,k} \cdot \frac{q_{i,v,g,k}}{\lambda_{m,d,g,k}} + s_{v,g,k} \cdot \frac{q_{i,v,g,k}}{\lambda_{v,a,g,k}} \right) \text{ e}$$

$$\frac{\partial \mathcal{L}}{\partial p_{i,d}} = \sum_g \sum_{k=0}^{n_g} \left(t_{g,k} \cdot \frac{q_{i,v,g,k} \cdot \lambda_{m,a,g,k}}{(\lambda_{v,d,g,k})^2} - s_{m,g,k} \cdot \frac{q_{i,v,g,k}}{\lambda_{v,d,g,k}} + t_{g,k} \cdot \frac{q_{i,m,g,k} \cdot \lambda_{v,a,g,k}}{(\lambda_{m,d,g,k})^2} - s_{v,g,k} \cdot \frac{q_{i,m,g,k}}{\lambda_{m,d,g,k}} \right),$$

que, fatorando pela quantidade de aparições do jogador no intervalo de jogo, pode ser

reescrito como em 4.6.

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial p_{i,a}} &= \sum_g \sum_{k=0}^{n_g} \left(q_{i,m,g,k} \left(-\frac{t_{g,k}}{\lambda_{v,d,g,k}} + \frac{s_{m,g,k}}{\lambda_{m,a,g,k}} \right) + q_{i,v,g,k} \left(-\frac{t_{g,k}}{\lambda_{m,d,g,k}} + \frac{s_{v,g,k}}{\lambda_{v,a,g,k}} \right) \right) \text{ e} \\
\frac{\partial \mathcal{L}}{\partial p_{i,d}} &= \sum_g \sum_{k=0}^{n_g} \left(q_{i,m,g,k} \left(\frac{t_{g,k}}{\lambda_{m,d,g,k}} \cdot \frac{\lambda_{v,a,g,k}}{\lambda_{m,d,g,k}} - \frac{s_{v,g,k}}{\lambda_{m,d,g,k}} \right) + q_{i,v,g,k} \left(\frac{t_{g,k}}{\lambda_{v,d,g,k}} \cdot \frac{\lambda_{m,a,g,k}}{\lambda_{v,d,g,k}} - \frac{s_{m,g,k}}{\lambda_{v,d,g,k}} \right) \right).
\end{aligned} \tag{4.6}$$

De onde podemos montar o vetor gradiente $\nabla \mathcal{L}$ como

$$\nabla_i \mathcal{L} = \begin{cases} \dfrac{\partial \mathcal{L}}{\partial p_{i,a}}, & \text{se } i \leq n_p \\ \dfrac{\partial \mathcal{L}}{\partial p_{i-n_p,d}}, & \text{c.c.} \end{cases}.$$

Além do gradiente, podemos buscar a hessiana de $\mathcal{L}$ e analisar a convexidade da função por meio da mesma, a qual é dada em 4.7.

$$\begin{aligned}
\frac{\partial^2 \mathcal{L}}{\partial p_{i,a} \partial p_{j,a}} &= \sum_g \sum_{k=0}^{n_g} \left(-q_{i,m,g,k} \cdot q_{j,m,g,k} \cdot \frac{s_{m,g,k}}{\lambda_{m,a,g,k}^2} - q_{i,v,g,k} \cdot q_{j,v,g,k} \cdot \frac{s_{v,g,k}}{\lambda_{v,a,g,k}^2} \right), \\
\frac{\partial^2 \mathcal{L}}{\partial p_{i,d} \partial p_{j,a}} &= \sum_g \sum_{k=0}^{n_g} \left(q_{i,m,g,k} \cdot q_{j,v,g,k} \cdot \frac{t_{g,k}}{\lambda_{m,d,g,k}^2} + q_{i,v,g,k} \cdot q_{j,m,g,k} \cdot \frac{t_{g,k}}{\lambda_{v,d,g,k}^2} \right), \\
\frac{\partial^2 \mathcal{L}}{\partial p_{i,a} \partial p_{j,d}} &= \sum_g \sum_{k=0}^{n_g} \left(q_{i,m,g,k} \cdot q_{j,v,g,k} \cdot \frac{t_{g,k}}{\lambda_{v,d,g,k}^2} + q_{i,v,g,k} \cdot q_{j,m,g,k} \cdot \frac{t_{g,k}}{\lambda_{m,d,g,k}^2} \right) \text{ e} \\
\frac{\partial^2 \mathcal{L}}{\partial p_{i,d} \partial p_{j,d}} &= \sum_g \sum_{k=0}^{n_g} \left(q_{i,m,g,k} \cdot q_{j,m,g,k} \cdot \left(\frac{s_{v,g,k}}{\lambda_{m,d,g,k}^2} - \frac{2 \cdot t_{g,k} \cdot \lambda_{v,a,g,k}}{\lambda_{m,d,g,k}^3} \right) \right) \\
&\quad + \sum_g \sum_{k=0}^{n_g} \left(q_{i,v,g,k} \cdot q_{j,v,g,k} \cdot \left(\frac{s_{m,g,k}}{\lambda_{v,d,g,k}^2} - \frac{2 \cdot t_{g,k} \cdot \lambda_{m,a,g,k}}{\lambda_{v,d,g,k}^3} \right) \right).
\end{aligned} \tag{4.7}$$

De modo análogo ao gradiente, podemos definir a matriz hessiana $\nabla^2 \mathcal{L}$ de forma matricial por meio das expressões de 4.7.

$$\nabla_{i,j}^2 \mathcal{L} = \begin{cases} \dfrac{\partial^2 \mathcal{L}}{\partial p_{i,a} \partial p_{j,a}}, & \text{se } i \leq n_p \text{ e } j \leq n_p \\ \dfrac{\partial^2 \mathcal{L}}{\partial p_{i,a} \partial p_{j-n_p,d}}, & \text{se } i \leq n_p \text{ e } j > n_p \\ \dfrac{\partial^2 \mathcal{L}}{\partial p_{i-n_p,d} \partial p_{j,a}}, & \text{se } i > \text{ e } j \leq n_p \\ \dfrac{\partial^2 \mathcal{L}}{\partial p_{i-n_p,d} \partial p_{j-n_p,d}}, & \text{se } i > \text{ e } j > n_p \end{cases}.$$

Por fim, para verificar a convexidade de $\mathcal{L}$ devemos ter que $\nabla^2 \mathcal{L}$ deve ser semi-definida positiva. Mas isso não é verdadeiro, visto que tomando qualquer vetor canônico e_l, temos que $e_l^\top \cdot \nabla^2 \mathcal{L} \cdot e_l = \nabla_{l,l}^2 \mathcal{L}$. Agora, escolhendo $l \leq n_p$, temos que

$$\begin{aligned}
\nabla_{l,l}^2 \mathcal{L} &= \sum_g \sum_{k=0}^{n_g} \left(-q_{l,m,g,k}^2 \cdot \frac{s_{m,g,k}}{\lambda_{m,a,g,k}^2} - q_{l,v,g,k}^2 \cdot \frac{s_{v,g,k}}{\lambda_{v,a,g,k}^2} \right) \\
&= -\sum_g \sum_{k=0}^{n_g} \left(q_{l,m,g,k}^2 \cdot \frac{s_{m,g,k}}{\lambda_{m,a,g,k}^2} + q_{l,v,g,k}^2 \cdot \frac{s_{v,g,k}}{\lambda_{v,a,g,k}^2} \right),
\end{aligned}$$

é um número negativo ou nulo, visto que os termos quadráticos são não negativos e os valores de $s_{m,g,k}$ e $s_{v,g,k}$ são correspondentes ao placar do jogo, logo também são não negativos.

Dessa forma, se escolhermos l de modo que esse vetor corresponda a direção de um jogador que entrou em campo em, pelo menos, uma subpartida que não terminou zero a zero temos que $e_l^\top \cdot \nabla^2\mathcal{L} \cdot e_l < 0$, o que mostra que $\nabla^2\mathcal{L}$ não é semi-definida positiva e, consequentemente, que $\mathcal{L}$ não é convexa.

Apesar da não convexidade de $\mathcal{L}$, foi elaborado um código em Python que, por meio da biblioteca `scipy`, busca os parâmetros que otimizam a função de verossimilhança. Foram testados diversos algoritmos para otimização, com e sem a utilização do gradiente e da hessiana, além de se valer da clusterização, onde os jogadores com até 20 subpartidas foram agrupados. Entretanto, em virtude da não convexidade e do espaço de parâmetros, que ainda possui uma dimensão alta, a otimização ainda se mostrou complexa. Diante do exposto, optou-se por deixar a abordagem frequentista de lado e se voltar a abordagem bayesiana.

4.4 Abordagem Bayesiana

Alternativamente a abordagem frequentista, apresentada na seção anterior, podemos modelar o futebol através do paradigma bayesiano. Nessa abordagem cada jogador continua sendo definido por um conjunto de parâmetros, porém tais parâmetros não são mais valores reais fixos, mas sim variáveis aleatórias. Com isso em mente, o objetivo passa, da estimação dos parâmetros de cada jogador, para a estimação da distribuição de probabilidade de cada uma dessas variáveis que representam os jogadores. Para tanto, necessitamos de um modelo e também de uma hipótese, à priori, de como os parâmetros das distribuições que modelam as observações se comportam.

Seguindo essa linha de pensamento, podemos, primeiramente, definir um modelo.

4.4.1 ADM - Attack and Defense Model

Para a modelagem inicial dessa abordagem, iremos nos valer do modelo de Poisson descrito na seção anterior, isso é, onde os gols em uma subpartida são expressos pelas equações em 4.3. Já os valores de $\lambda_{m,a,g,k}$, $\lambda_{v,a,g,k}$, $\lambda_{m,d,g,k}$ e $\lambda_{v,d,g,k}$, os quais, na abordagem frequentista, foram tomados como em 4.5 (uma generalização de 4.2), são definidos de modo análogo, bastando definir se a clusterização, explicada e utilizada na seção anterior, será ou não aplicada a este modelo. Para essa modelagem, a mesma não foi utilizada, de modo que definimos tais valores conforme as Equações em 4.2. Novamente, devemos lembrar que não faz sentido termos valores negativos para as proficiências dos jogadores,

pois isso implicaria numa piora do desempenho do time com tal jogador em campo em relação a quando o time joga com um jogador a menos.

Com o modelo definido, passamos a elicitação da nossa hipótese à priori. Aqui, como cada jogador é dado por uma distribuição, é inviável definir, manualmente, priori por priori. Algumas possibilidades são plausíveis para elicitação das mesmas, como a utilização do valor de mercado dos jogadores ou do clube em que os mesmos atuam como forma de balancear os jogadores inicialmente.

Tais ideias partem das premissas de que, no caso da primeira, jogadores mais caros são, em geral, melhores. Para a segunda ideia, adicionamos a premissa de que os jogadores de um mesmo clube tendem a ter desempenhos similares. Assim, juntando com a premissa utilizada na primeira ideia, teríamos que a qualidade técnica de todos os jogadores de um time deve ser similar, logo, o valor de mercado do clube poderia ser a base da priori para todos os jogadores. Naturalmente, há exceções. Jogadores que jogam em mercados com menor visibilidade tendem a ter preços menores, uma vez que são pouco observados.

Essas prioris são razoáveis (os valores de mercado são sempre positivos) e podem ser utilizadas com facilidade por meio de dados do Transfermarkt (TRANSFERMARKT. . . , 2022). Entretanto, para o escopo desse trabalho, foi utilizada a mesma priori para todos os jogadores, de modo que, apesar da falta de informação inicial, facilitará na implementação do modelo. Nesse sentido, a priori utilizada foi uma distribuição semi-normal, isso é, a distribuição de $Y = |X|$, sendo $X \sim N(0, 1)$. Ou, por simplicidade de implementação, uma normal padrão truncada em zero.

Essa priori, além de poder ter a interpretabilidade de que os atletas possuem um nível de habilidade dado por uma normal, poderia facilitar algumas contas ao realizar os cálculos de modo analítico. Sabendo que a verossimilhança do modelo proposto é dada por uma Poisson, ter uma priori Gamma para o parâmetro da Poisson pode nos dar, diretamente, a posteriori. Como a priori dada é semi-normal, o parâmetro da Poisson tomada como verossimilhança será uma variável aleatória Chi-quadrado, a qual é um caso especial da Gamma, pois se dá pela razão de somas de variáveis semi-normais. Dessa forma, os cálculos analíticos podem ser simplificados.

Entretanto, dado o número elevado de jogos e jogadores a serem modelados, optou-se pelo uso do algoritmo HMC por meio do Stan (STAN DEVELOPMENT TEAM, 2018), de modo a obter as amostras das distribuições à posteriori dos parâmetros de interesse ao invés de calculá-los analiticamente. Com o modelo e a priori definida partiu-se para a implementação do modelo, o qual está disponível no Apêndice A.1.

Como entrada, o modelo recebe informações sobre o número de subpartidas que estão sendo modeladas, bem como o número de jogadores e o vetor de tempo de cada uma das subpartidas. Além disso, temos o número de jogadores por clube em cada uma das subpartidas, apenas para possibilitar a criação dos *arrays* para cada um dos clubes. Por

fim, os últimos três dados de entrada são referentes aos dados observados de fato, isso é, os placares de cada uma das subpartidas e cada uma das escalações (mandante e visitante), também para cada uma das subpartidas.

A definição dos parâmetros e do modelo leva em consideração o que desejamos amostrar. Como queremos amostrar as distribuições de $p_{i,a}$ e de $p_{i,d}$ para todos os possíveis jogadores i, os parâmetros do modelo se referem justamente a isso, onde cada um dos vetores corresponde a uma dessas forças. Como comentado previamente, tais parâmetros tem suporte nos reais positivos, o que também já foi restrito na definição dos parâmetros.

Por fim, o modelo deve receber uma priori para cada um dos parâmetros, a qual, como discutido anteriormente, é dada por uma semi-normal. Uma vez que os parâmetros já estão restritos a $\mathbb{R}_+$, podemos considerar que a priori é dada por uma normal padrão. Além da priori, o modelo deve possuir uma função de verossimilhança para que o mesmo possa realizar o cálculo da distribuição, a qual é dada conforme o modelo em 4.3.

4.4.2 HAM - Home Away Model

Agora, com um primeiro modelo definido, podemos voltar a Figura 5, onde temos os histogramas de gols de cada partida, discriminados pelo mando de campo. Apesar do teste de hipóteses realizado, cujos resultados foram apresentados na Tabela 1, falhar em rejeitar a hipótese de que o número de gols em uma partida de futebol é dada por uma distribuição de Poisson, não temos a informação acerca do número de gols do mandante e do visitante serem dados pela mesma distribuição. Podemos ver que a mesma figura nos mostra que as médias dos gols é diferente em relação ao mando. Dessa forma, podemos questionar se há diferença entre a distribuição de gols do mandante para o visitante, isso é, o mandante possui alguma vantagem ou desvantagem em relação ao visitante? Essa pergunta é equivalente a nos perguntarmos acerca da igualdade entre as médias das distribuições que estamos modelando.

A resposta para essa pergunta é incerta, mas podemos levantar a hipótese de que duas as médias são iguais, ou seja, nenhum dos times possui vantagem. Nesse cenário, teríamos $s_{m,g,k}, s_{v,g,k} \sim \text{Poisson}(\lambda)$, para algum λ. Para testar essa hipótese, podemos realizar um teste t de Student para média de duas amostras, já implementado na biblioteca `scipy`, a qual retorna a estatística de teste e o p-valor do teste. Para esse teste utilizaremos $\alpha = 0.01$, logo, se o p-valor retornado pela função ao receber as duas séries de gols for menor que α, rejeitamos a hipótese nula, o que pode ser interpretado como uma evidência de que, ao menos minimamente, os times são influenciados pelo mando de campo.

Realizando o teste encontramos um p-valor de, aproximadamente, $9.852 \cdot 10^{-51}$, o que nos faz rejeitar a hipótese de que as duas distribuições possuem médias iguais. Uma vez que tal hipótese foi rejeitada, podemos pensar em realizar algumas pequenas alterações no modelo de modo que possamos diferenciar o fator mando de campo. Duas modelagens

surgem naturalmente, uma em que cada jogador possui um rendimento diferente conforme o mando de campo, isso é, o jogador i tem rendimento X jogando em casa e Y jogando fora, ou onde cada clube possui um fator mando de campo, como se sua torcida fosse o décimo segundo jogador. Tais modelos podem ser definidos, matematicamente, por meio de pequenas alterações no modelo dado em 4.3.

4.4.2.1 HAM1

Partindo para a modelagem dos jogadores junto ao mando de campo podemos considerar, primeiramente, o modelo em que os jogadores possuem rendimentos distintos conforme o local da partida (se o jogo é em casa ou fora). Nesse sentido, o modelo proposto previamente pode servir como modelo base e, por meio de pequenas alterações, podemos chegar ao modelo desejado.

Como a diferença entre tais modelos se dá pelo rendimento dos atletas conforme o mando de campo, a mudança natural é na inclusão de um novo conjunto de parâmetros para os jogadores, onde discrimina-se o mando. Dessa forma, os parâmetros $p_{i,a}$ e $p_{i,d}$ dão lugar aos parâmetros p_{i,a_m}, p_{i,d_m}, p_{i,a_v} e p_{i,d_v}, os quais possuem a mesma interpretação dos parâmetros originais, mas incorporando a discriminação do mando. Com a alteração dos parâmetros dos jogadores, é natural que as forças dos clubes também se alterassem, de modo que, para esse modelo, os parâmetros dos clubes passam a ser expressos pelo conjunto de Equações 4.8.

$$\begin{aligned}
\lambda'_{m,a,g,k} &= \sum_{i \in \mathcal{C}_{m,g,k}} p_{i,a_m}, \\
\lambda'_{v,a,g,k} &= \sum_{i \in \mathcal{C}_{v,g,k}} p_{i,a_v}, \\
\lambda'_{m,d,g,k} &= \sum_{i \in \mathcal{C}_{m,g,k}} p_{i,d_m} \text{ e} \\
\lambda'_{v,d,g,k} &= \sum_{i \in \mathcal{C}_{v,g,k}} p_{i,d_v},
\end{aligned} \tag{4.8}$$

Finalmente, o modelo final possui a mesma forma do modelo anterior, sendo dado por 4.9.

$$\begin{aligned}
s_{m,g,k} &\sim \text{Poisson}\left(\frac{\lambda'_{m,a,g,k}}{\lambda'_{v,d,g,k}} \cdot t_{g,k}\right) \\
s_{v,g,k} &\sim \text{Poisson}\left(\frac{\lambda'_{v,a,g,k}}{\lambda'_{m,d,g,k}} \cdot t_{g,k}\right),
\end{aligned} \tag{4.9}$$

Novamente, utilizando uma priori semi-normal para os parâmetros dos jogadores, podemos elaborar o código Stan para tal modelo. O código, assim como no modelo anterior, está disponível no Apêndice A.2. Como o modelo propriamente dito, a implementação também é bem similar a anterior, havendo apenas a alteração dos parâmetros e, consequentemente, da média da Poisson.

4.4.2.2 HAM2

Por fim, além de modelar a diferença que o mando de campo gera por meio dos jogadores, pode-se realizar tal modelagem de modo que essa diferença seja explicada por meio do clube mandante. Nesse caso, cada clube possui um parâmetro que diferencia sua atuação em casa para a atuação fora de casa, servindo, literalmente, como um décimo segundo jogador. Desse modo, o mando de campo terá influência para melhorar tanto o ataque quanto a defesa do clube mandante. Matematicamente, podemos traduzir essa ideia por meio do modelo em 4.10.

$$\begin{aligned} s_{m,g,k} &\sim \text{Poisson}\left(\left(\frac{\lambda_{m,a,g,k} + \kappa_{m,a,g,k}}{\lambda_{v,d,g,k}}\right) \cdot t_{g,k}\right) \\ s_{v,g,k} &\sim \text{Poisson}\left(\left(\frac{\lambda_{v,a,g,k}}{\lambda_{m,d,g,k} + \kappa_{m,d,g,k}}\right) \cdot t_{g,k}\right). \end{aligned} \tag{4.10}$$

Os valores de $\lambda_{m,a,g,k}$, $\lambda_{v,a,g,k}$, $\lambda_{m,d,g,k}$ e $\lambda_{v,d,g,k}$ foram tomados como em 4.5, enquanto $\kappa_{m,a,g,k}$ e $\kappa_{m,d,g,k}$ são novos parâmetros, também com suporte em $\mathbb{R}_+$.

Por meio do Modelo 4.10, podemos ver que o mando de campo se encaixa exatamente como mais um jogador para o clube mandante, o que nos permite ver com maior clareza a possibilidade que esse fator tem de neutralizar a equipe adversária. Entretanto, também por meio dessas equações, pode-se ver que a efetividade do mando de campo também depende do adversário, isso é, o mando de campo por si só não garante que o clube mandante irá ter um determinado resultado.

Novamente, tal modelo foi implementado no Stan tendo como base o código inicial, do modelo ADM. O código final para esse modelo pode ser visto no Apêndice A.3.

Para tal modelo, além dos dados que o modelo ADM utiliza, necessitamos inserir o número de clubes que estão sendo modelados e um *array* mapeando o clube mandante de cada uma das partidas. As demais alterações no código são referentes aos parâmetros que serão estimados, os quais foram incluídos no modelo. Mais uma vez, atribuiu-se uma priori de distribuição semi-normal aos mesmos. Por fim, foram realizadas alterações na média da Poisson para que a mesma correspondesse ao modelo desejado.

Por fim, vale lembrar que tais modelos sofrem com um problema de identificabilidade, ou seja, assim como visto na Seção 4.3, multiplicar ou dividir todos os parâmetros não irá alterar os modelos.

4.4.3 Rodando os modelos

Após definir os modelos podemos realizar a execução dos mesmos, obtendo amostras das distribuições a posteriori dos parâmetros dos jogadores. Devemos lembrar que o número de jogadores modelados, bem como o de subpartidas, é muito grande, logo tais execuções são custosas. Além disso, por se tratar de algoritmos que usam cadeias de Markov, as

paralelizações possíveis são na execução simultânea de diferentes cadeias para cada um dos modelos, entretanto isso também detém um custo computacional. Nesse sentido, com o intuito de otimizar a execução, driblando o custo computacional, utilizou-se, assim como no processo de raspagem, o GitHub Actions para executar o modelo.

4.4.3.1 Resultados

Tendo rodado os modelos, podemos partir para a análise dos resultados. Os principais resultados serão acerca dos melhores jogadores de modo geral. Para tanto, podemos ranquear os jogadores por meio do produto de suas proficiências, assim, quanto maior o produto, melhor sua posição no ranking. Essa métrica, derivada da média geométrica, auxilia a traduzir o impacto de cada jogador em uma subpartida. Tal impacto pode ser visto por meio da média da distribuição de gols. Para o número de gols do seu clube atacante, a variação da proficiência de ataque de um jogador, aumentando sua proficiência, influencia positivamente a média, fazendo com que a mesma aumente. Por outro lado, em relação ao número de gols do adversário, o aumento da proficiência de defesa de um jogador faz com que a média de gols do seu adversário diminua, uma vez que esse parâmetro aparece no denominador.

Para elaboração desse ranking, a proficiência de ataque e de defesa utilizada será a proficiência média da distribuição amostrada para cada jogador. Em relação ao modelo HAM2, o qual possui duas forças de ataque e duas de defesa para cada jogador, as proficiências utilizadas se dão pela média de todas as proficiências de ataque e pela média de todas as proficiências de defesa.

Os melhores jogadores do período de 2018 a 2022, segundo essa métrica, podem ser vistos na Tabela 4.

Pos.	ADM	HAM1	HAM2
1	Sávio (VIL/GO)	Gustavo (OES/SP)	Vitinho (GUA/SP)
2	Bruno (NOV/SP)	Tiago C. (BOT/SP)	Maicon (BRA/RS)
3	Vitinho (GUA/SP)	Marlon (OES/SP)	Bruno (NOV/SP)
4	Henrique (CSA/AL)	Werik (RBB/SP)	Rincon (CUI/MT)
5	Maicon (BRA/RS)	Caue (JUV/RS)	Sávio (VIL/GO)
6	Leandro (OPE/PR)	Landazuri (FOR/CE)	Leandro (OPE/PR)
7	Fernando (LON/PR)	Leo Bahia (BRA/RS)	Fernando (LON/PR)
8	Lessa (BOT/RJ)	Leonardo (INT/RS)	Sávio (CAM/MG)
9	Sávio (CAM/MG)	Bruno (GUA/SP)	Lessa (BOT/RJ)
10	Darli (CRB/AL)	Vladimir (AVA/SC)	Darli (CRB/AL)

Tabela 4 – Ranking: Melhores jogadores segundo cada modelo.

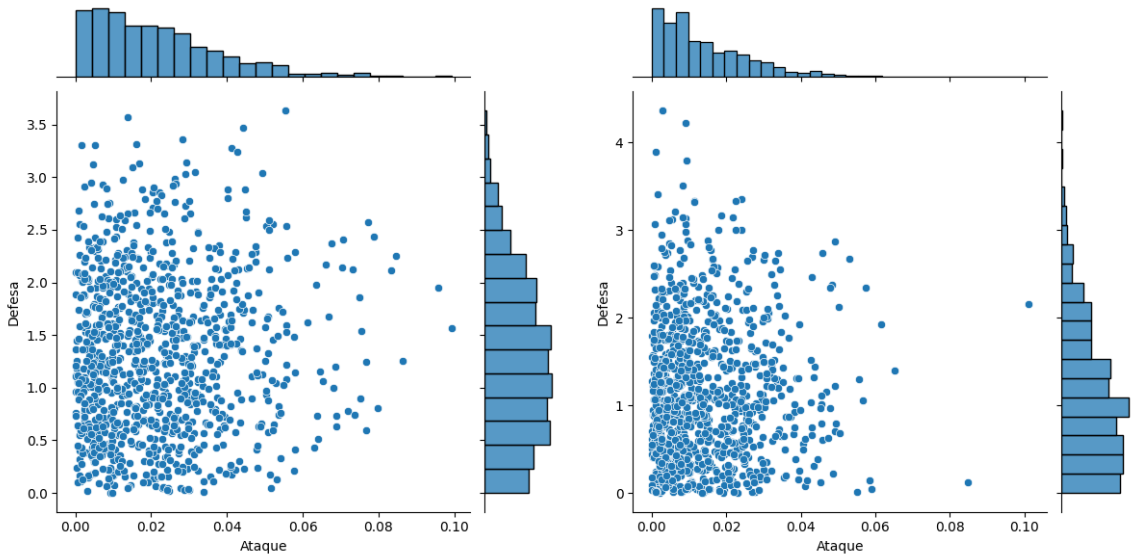
Note que os rankings dos modelos ADM e HAM2 são bem parecidos, com apenas um jogador diferente, exceto pela ordem. Isso é razoável, uma vez que os modelos são similares, se alterando apenas por meio do fator mando de campo, o qual, no modelo HAM2,

atua como um jogador a mais. Além disso, muitos desses jogadores atuam em clubes de menor expressão, como Sávio, no Vila Nova, e Vitinho, no Guarani, onde possivelmente também devem ter atuado pouco. Uma forma de lidar com esse problema é filtrar os jogadores, de modo que só entrem para o ranking os atletas que participaram de, ao menos, 100 subpartidas. Nesse sentido, podemos elaborar o ranking visto na Tabela 5.

Pos.	ADM	HAM1	HAM2
1	Wellington (ATL/GO)	Juninho (AME/MG)	Wellington (ATL/GO)
2	Ricardo (CRB/AL)	Sornoza (COR/SP)	Giuliano (COR/SP)
3	Wellington (AME/MG)	Jarro (BRA/RS)	Ricardo (CRB/AL)
4	Giuliano (COR/SP)	Léo Artur (SAM/MA)	Daniel (BOT/RJ)
5	Igor Carius (CUI/MT)	Ramon (ATL/GO)	Wellington (AME/MG)
6	Daniel (BOT/RJ)	Rayan (CRI/SC)	Igor Carius (CUI/MT)
7	Sornoza (COR/SP)	Matheus (CEA/CE)	Sornoza (COR/SP)
8	Mansur (SAO/SP)	Roni (COR/SP)	Mansur (SAO/SP)
9	Leo (CHA/SC)	Wellington (AME/MG)	Augusto (LON/PR)
10	Tadeu (GOI/GO)	Tadeu (GOI/GO)	Leo (CHA/SC)

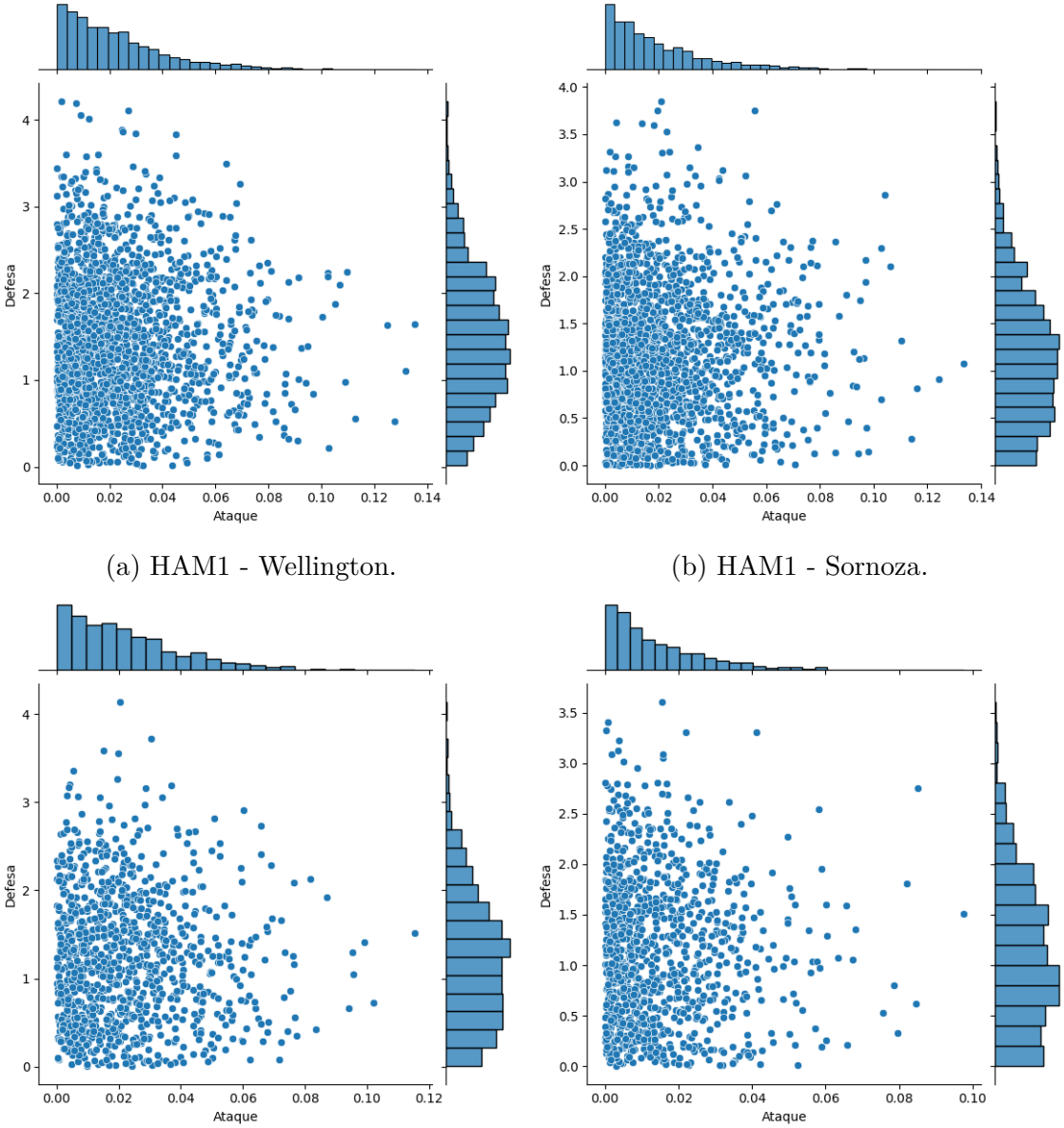
Tabela 5 – Ranking: Melhores jogadores, com pelo menos 100 subpartidas disputadas, segundo cada modelo.

Podemos ver que, novamente, os rankings do ADM e do HAM2 são bem similares. Além disso, já aparecem mais jogadores conhecidos. Dois jogadores que chamam atenção nesse ranking são o Wellington, do América Mineiro, e Sornoza, ex-jogador do Corinthians, os quais aparecem no ranking de todos os modelos. Dado isso, podemos ver como sua distribuição se comporta, iteração a iteração da Cadeia de Markov, conforme Figuras 7 e 8.



(a) AMD - Wellington. (b) AMD - Sornoza.

Figura 7 – Distribuições ADM - Wellington e Sornoza.



(a) HAM1 - Wellington.

(b) HAM1 - Sornoza.

(c) HAM2 - Wellington.

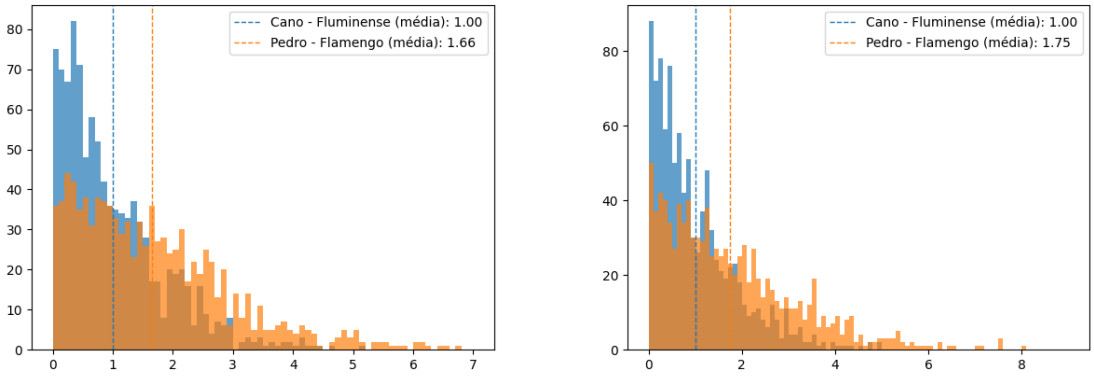
(d) HAM2 - Sornoza.

Figura 8 – Distribuições HAM - Wellington e Sornoza.

Pode-se ver que as distribuições são similares. Além disso, os parâmetros defensivos possuem valores muito mais elevados que os parâmetros ofensivos, o que já era esperado. Visto que, em todos os modelos, a média da distribuição de Poisson é dada pelo produto entre a razão entre as forças das equipes e o tempo de jogo, podemos interpretar tal razão como a média de gols em um minuto. Como o número de gols esperados em um minuto dentro de um jogo de futebol é muito baixo, faz sentido que os parâmetros defensivos sejam maiores que os ofensivos.

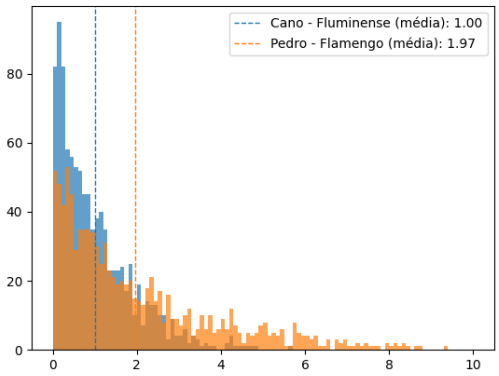
Outra análise interessante é a de comparar qual jogador é melhor. Nesse sentido pode-se comparar, por exemplo, dois atacantes da dupla Fla-Flu: Pedro e Germán Cano. Como ambos são jogadores referência no ataque, atuando como centroavante, seus parâmetros de defesa podem ser desconsiderados para a avaliação, uma vez que a ideia é ver quem

contribui mais para fazer gols. Dessa forma, podemos ilustrar, graficamente, a distribuição dos parâmetros ofensivos de cada um dos dois atletas. Novamente, devemos lembrar do problema de identificabilidade do modelo, logo podemos fazer uma padronização em tais parâmetros, de modo que a média do parâmetro ofensivo do atacante tricolor seja fixada em 1 com o intuito de melhorar a escala para a visualização. No caso do HAM1, a normalização se dá, apenas, na proficiência de ataque em casa.

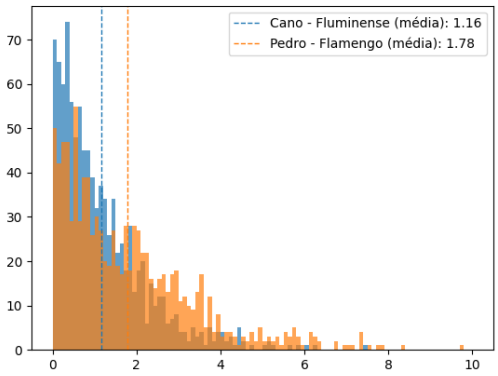


(a) AMD.

(b) HAM2.



(c) HAM1 - casa.



(d) HAM1 - fora.

Figura 9 – Distribuições ofensivas: Germán Cano e Pedro no período de 2018 a 2022.

Conforme Figura 9, pode-se ver que, segundo o HAM1, jogando em casa Pedro é, em média, quase duas vezes mais eficiente que Germán Cano. Por outro lado, quando estão fora de seus domínios, Germán Cano é um pouco mais efetivo, enquanto Pedro diminui seu rendimento, o que faz a diferença entre eles diminuir um pouco, mas ainda com vantagem considerável para o atacante flamenguista. Em relação aos outros dois modelos, Pedro também se mostra mais efetivo, com margens, em relação a média, similares.

Por fim, além do ranking de jogadores, o HAM2 nos possibilita ranquear os clubes. Tal ranking pode ser elaborado de modo análogo ao dos jogadores, por meio da comparação dos produtos entre a média da distribuição ofensiva e defensiva que, no caso dos clubes,

podem ser interpretados como o quanto a torcida (ou algum outro fator) influencia apoiando o clube, tanto no setor ofensivo quanto no defensivo. Dessa forma, considerando o período de 2018 a 2022, os clubes com o maior impacto podem ser encontrados na Tabela 6.

Pos.	Clube
1	Joinville / SC
2	Icasa / CE
3	Mogi Mirim / SP
4	Vila Nova / GO
5	Athletico Paranaense / PR
6	Avaí / SC
7	ABC / RN
8	CRB / AL
9	São Caetano / SP
10	Paysandu / PA

Tabela 6 – Ranking dos clube com maior influência do mando de campo.

Como pode-se ver, alguns clubes de menor expressão, predominantemente clubes que jogaram a série B nesse período, tomam conta das primeiras posições do ranking. Tal posicionamento faz sentido, uma vez que, muitas vezes, clubes da série B possuem um orçamento menor, além de maiores problemas logísticos o que pode acarretar na piora do seu desempenho nas partidas que joga como visitante. Consequentemente, o desempenho como mandante do seu adversário tende a melhorar, fato que reflete no parâmetro κ desses clubes.

Um clube que vai à contramão da ideia apresentada no parágrafo anterior é o Athletico Paranaense, que figura na quinta posição. Novamente, sua posição tem uma explicação plausível: o Athletico, até 2020, era o único clube que utilizava a grama sintética. Logo, jogar na Arena da Baixada era, muitas vezes um desafio um pouco maior para os adversários, acostumados com o gramado natural. De modo análogo, o Athletico enfrentava um desafio um pouco fora do habitual ao jogar fora de casa, o que pode explicar sua aparição nesse ranking.

5 Conclusão

A estatística é uma das principais ferramentas para análise de dados e inferência. No presente trabalho analisamos sua aplicação no contexto do futebol brasileiro, uma vez que, com tais análises, pode-se ter um melhor embasamento durante discussões acerca o futebol, além de, é claro, no contexto das casas de apostas, onde cada informação a mais pode ser relevante, tanto para o apostador quanto para a banca.

Para tanto, o trabalho se iniciou com a raspagem e tratamento dos dados disponíveis nas súmulas da CBF, processo esse que acabou por ser automatizado com o intuito de obter os dados conforme os mesmos fossem disponibilizados.

Após a raspagem dos dados iniciou-se o processo de modelagem, no qual abordamos o problema tanto no paradigma frequentista quanto no bayesiano. Infelizmente a abordagem frequentista não funcionou conforme o esperado, uma vez que a otimização da função de verossimilhança se mostrou extremamente custosa. Desse modo tal abordagem foi deixada de lado.

Por outro lado, a abordagem bayesiana foi implementada, com três modelos propostos, e executada sem grandes problemas. O primeiro modelo modela os gols da partida simplesmente por meio dos jogadores que estavam em campo no momento do gol, com cada jogador tendo uma proficiência de ataque e outra de defesa. Já o segundo modelo discrimina o mando de campo, isso é, atribuí diferentes parâmetros para os jogadores conforme o local da partida, o que fez com que cada atleta possuísse quatro parâmetros. Por fim, o último modelo buscou tratar a questão do mando de campo de modo a não inflar muito o número de parâmetros, assim tal modelo interpretou o clube mandante como um novo jogador, o qual atua a favor do mandante da partida, como uma partida 12 contra 11.

Ainda assim, em virtude do alto número de parâmetros a serem estimados, a complexidade do programa ficou alta. Para contornar o problema computacional o código foi rodado nos servidores do GitHub o que possibilitou, além do maior poder computacional de tais servidores, a paralelização de alguns dos processos, fazendo com que cada servidor executasse algumas cadeias de Markov em paralelo.

Essa abordagem acabou gerando resultados interessantes, como o ranking dos melhores jogadores e a comparação de dois atacantes que foram referência no Brasil em 2022: Pedro e Germán Cano, onde, além de comparar, vimos a diferença entre o desempenho de ambos jogando em casa ou fora de casa. Por fim, também foi analisado o comportamento do fator mando de campo por meio do ranking das equipes que mais se valem de tal fator.

6 Trabalhos Futuros

Por se tratar de um trabalho envolvendo extração de dados, o número de possibilidades de expansão é grade. Entre elas, pode-se citar a análise de redes voltadas aos clubes e jogadores, como a análise de quais clubes mais realizam transferência de jogadores entre si, ou a clusterização de jogadores, por meio de uma rede onde os jogadores são nós e as arestas são ponderadas de acordo com o número de partidas (ou minutos) que cada jogador passou em campo tendo outro jogador como companheiro de time, por exemplo.

Além dos projetos envolvendo redes, outras possibilidades são a aplicação de diferentes modelos nos dados, seja para modelar os jogadores ou para predição, valendo ressaltar que o trabalho de predição necessita de alguma heurística para definir as escalações de cada time para previsões de longo prazo. No caso do modelo frequentista essa modelagem pode ser realizada por meio da correção ou reimplementação do modelo aqui apresentado, talvez com alguma outra clusterização ou uso de outro otimizador. Já em relação a abordagem bayesiana, a mesma pode ser variada por meio da modelagem dos gols ou da priori utilizada. Ao variar a priori utilizada, um trabalho interessante pode ser a tentativa de precificar os jogadores, através da utilização do valor de mercado como priori.

Ainda no âmbito de previsões, uma possibilidade é a extensão do processo de raspagem para possibilitar a coleta das informações sobre os técnicos, possibilitando utilizar essa informação na modelagem. Tal ideia se dá pela grande mudança de postura observada recentemente em algumas equipes, como no Flamengo, onde a mudança de técnico, no ano de 2022, foi determinante na forma de jogar da equipe, saindo de uma equipe próxima do rebaixamento para a conquista da Copa do Brasil e da Copa Libertadores.

Referências

BEAUDOIN, David; SWARTZ, Tim. Strategies for Pulling the Goalie in Hockey. **The American Statistician**, v. 64, p. 197–204, ago. 2010. DOI: 10.1198/tast.2010.09147.

CBF. **Confederação Brasileira de Futebol**. Disponível em: <https://www.cbf.com.br>. Acesso em: 17 out. 2022.

FERREIRA, Bernardo Russo Guedes. **Previsão de resultados de tênis**. [S.l.: s.n.], 2018. Disponível em: <https://hdl.handle.net/10438/27857>.

GARCIA, Pedro. **Apostas esportivas online movimentam R$ 4 bilhões por ano no Brasil**. Disponível em: <https://www.gaz.com.br/apostas-esportivas-online-movimentam-r-4-bilhoes-por-ano-no-brasil/>. Acesso em: 15 out. 2022.

GVR. **Sports Betting Market Size, Share & Trends Analysis By Platform, By Type, By Sports Type (Football, Basketball, Baseball, Horse Racing, Cricket, Hockey, Others), By Region, And Segment Forecasts, 2022 - 2030**. Disponível em: <https://www.grandviewresearch.com/industry-analysis/sports-betting-market-report>. Acesso em: 17 out. 2022.

HISTÓRIA DO FUTEBOL NO BRASIL. **História do Futebol no Brasil - Wikipedia, The Free Encyclopedia**. Disponível em: <https://pt.wikipedia.org/wiki/Hist%C3%B3ria_do_futebol_do_Brasil>. Acesso em: 15 out. 2022.

JONES, Eric Scot. **PREDICTING OUTCOMES OF NBA BASKETBALL GAMES**. [S.l.: s.n.], 2016. Disponível em: <https://library.ndsu.edu/ir/bitstream/handle/10365/28084/Predicting%5C%20Outcomes%5C%20of%5C%20NBA%5C%20Basketball%5C%20Games.pdf?sequence=1&isAllowed=y>.

KARLIS, Dimitris; NTZOUFRAS, Ioannis. Analysis of Sports Data by Using Bivariate Poisson Models. **Journal of the Royal Statistical Society. Series D (The Statistician)**, [Royal Statistical Society, Wiley], v. 52, n. 3, p. 381–393, 2003. ISSN 00390526, 14679884. Disponível em: <http://www.jstor.org/stable/4128211>. Acesso em: 18 out. 2022.

_____. On modeling soccer data. **Student**, v. 3, n. 4, p. 229–244, 2000.

LEE, Alan. Modeling Scores in the Premier League: Is Manchester United Really the Best? **CHANCE**, v. 10, p. 15–19, set. 2012. DOI: 10.1080/09332480.1997.10554791.

MAIA, Luiz Fernando Guilhem Nassif. **Modelos in-play para partidas do Campeonato Brasileiro de Futebol**. [S.l.: s.n.], 2021. Disponível em: <https://hdl.handle.net/10438/31404>.

NIST/SEMATECH. **e-Handbook of Statistical Methods**. 2012. Disponível em: <https://www.itl.nist.gov/div898/handbook/eda/section3/eda35f.htm>. Acesso em: 20 out. 2022.

O ECONOMISTA. **Futebol e economia sempre caminham lado a lado**. Disponível em: <https://www.oeconomista.com.br/futebol-e-economia-caminham-lado-a-lado/>. Acesso em: 15 out. 2022.

SANTOS, João Marcos Amorim dos. **Previsões de resultados em partidas do campeonato brasileiro de futebol**. [S.l.: s.n.], 2019. Disponível em: <https://hdl.handle.net/10438/27672>.

STAN DEVELOPMENT TEAM. **The Stan Core Library**. [S.l.: s.n.], 2018. Version 2.18.0. Disponível em: <http://mc-stan.org/%209>.

TRANSFERMARKT. Disponível em: <https://www.transfermarkt.com.br/>. Acesso em: 16 nov. 2022.

WOLF, Stephan; SCHMITT, Maximilian; SCHULLER, Bjorn. A football player rating system. **Journal of Sports Analytics**, p. 243–257, 2020. DOI: 10.3233/JSA-200411.

Apêndices

APÊNDICE A – Códigos Stan

A.1 Attack and Defense Model

Código Stan para o modelo visto na Seção 4.4.1.

```
data {
  int<lower = 1> n_obs;
  int<lower = 1> n_players;
  array[n_obs] int<lower = 1> times;
  int<lower = 1> n_players_per_game;
  array[n_obs, 2] int results;
  array[n_obs, n_players_per_game] int club_1;
  array[n_obs, n_players_per_game] int club_2;
}

parameters {
  array[n_players] real<lower = 0> theta_atk;
  array[n_players] real<lower = 0> theta_def;
}

model {
  theta_atk ~ std_normal();
  theta_def ~ std_normal();
  for (n in 1:n_obs){
    results[n, 1] ~ poisson(sum(theta_atk[club_1[n, ]]) /
                            sum(theta_def[club_2[n, ]]) *
                            times[n]);
    results[n, 2] ~ poisson(sum(theta_atk[club_2[n, ]]) /
                            sum(theta_def[club_1[n, ]]) *
                            times[n]);
  }
}
```

A.2 Home Away Model 1

Código Stan para o modelo visto na Seção 4.4.2.1.

```
data {
  int<lower = 1> n_obs;
  int<lower = 1> n_players;
  array[n_obs] int<lower = 1> times;
  int<lower = 1> n_players_per_game;
  array[n_obs, 2] int results;
  array[n_obs, n_players_per_game] int club_1;
  array[n_obs, n_players_per_game] int club_2;
}

parameters {
  array[n_players] real<lower = 0> theta_atk_m;
  array[n_players] real<lower = 0> theta_def_m;
  array[n_players] real<lower = 0> theta_atk_v;
  array[n_players] real<lower = 0> theta_def_v;
}

model {
  theta_atk_m ~ std_normal();
  theta_def_m ~ std_normal();
  theta_atk_v ~ std_normal();
  theta_def_v ~ std_normal();
  for (n in 1:n_obs){
    results[n, 1] ~ poisson(sum(theta_atk_m[club_1[n, ]]) /
                    sum(theta_def_v[club_2[n, ]]) *
                    times[n]);
    results[n, 2] ~ poisson(sum(theta_atk_v[club_2[n, ]]) /
                    sum(theta_def_m[club_1[n, ]]) *
                    times[n]);
  }
}
```

A.3 Home Away Model 2

Código Stan para o modelo visto na Seção 4.4.2.2.

```
data {
  int n_obs;
  int n_clubs;
  int n_players;
```

```
  int n_players_per_game;
  array[n_obs] int times;
  array[n_obs] int home_clubs;
  array[n_obs, 2] int results;
  array[n_obs, n_players_per_game] int club_1;
  array[n_obs, n_players_per_game] int club_2;
}

parameters {
  array[n_players] real<lower = 0> theta_atk;
  array[n_players] real<lower = 0> theta_def;
  array[n_clubs] real<lower = 0> sigma_atk;
  array[n_clubs] real<lower = 0> sigma_def;
}

model {
  theta_atk ~ std_normal();
  theta_def ~ std_normal();
  sigma_atk ~ std_normal();
  sigma_def ~ std_normal();
  for (n in 1:n_obs){
    results[n, 1] ~ poisson(((sum(theta_atk[club_1[n, ]]) +
                               sigma_atk[home_clubs[n]]) /
                              sum(theta_def[club_2[n, ]])) *
                             times[n]);
    results[n, 2] ~ poisson((sum(theta_atk[club_2[n, ]]) /
                              (sum(theta_def[club_1[n, ]]) +
                               sigma_def[home_clubs[n]])) *
                             times[n]);
  }
}
```