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

Previsão Probabilística dos Resultados da Copa do Mundo de 2022 usando uma abordagem Bayesiana

Rodrigo Hideki Tozaki Moribayashi
Orientador: Luis Ernesto Bueno Salasar
Coorientador: Marcio Alves Diniz

Trabalho de Graduação apresentado como parte dos requisitos para obtenção do título de Bacharel em Estatística.

São Carlos
Setembro de 2022

Resumo

Este trabalho tem como objetivo construir e avaliar um modelo estatístico preditivo para jogos de futebol. Para isso, consideramos uma abordagem bayesiana em um novo modelo inspirado no de Lee (1997). A inspiração vem de dois pontos citados pelo autor: o fato dele considerar os números de gols marcados pelos dois times como variáveis independentes e com distribuição de Poisson, e que o número médio de gols dos times uma partida são explicados pelos efeitos ofensivos e defensivos das equipes. Neste trabalho, propomos modelar esses efeitos para as seleções participantes da Copa do Mundo de 2022, por meio da pontuação dos jogadores no jogo eletrônico FIFA 23. Esses efeitos são considerados conjuntamente às *odds* de sites de aposta esportiva, estas representadas por parâmetros de uma priori com distribuição Gamma. O procedimento bayesiano é aplicado para obtenção das distribuições marginais dos dados (números de gols de cada time), permitindo assim o cálculo das probabilidades associadas às partidas.

Palavras-chave: *futebol, Copa do Mundo, previsão, bayesiana, FIFA, odds, modelagem.*

Conteúdo

Capítulo 1

Introdução

A Copa do Mundo é o maior evento esportivo do planeta. Realizada a cada quatro anos, é disputada por 32 times que representam seus respectivos países. A primeira etapa da competição é uma fase de grupos, com quatro times por grupo. Cada time joga uma vez com cada integrante de seu grupo, fazendo três jogos, o que leva a um total de 48 partidas nessa etapa do torneio. Os dois melhores classificados de cada grupo avançam para a fase final do torneio, em que confrontos únicos de caráter eliminatório definem os times que seguem no torneio, até que reste somente um time, que é declarado campeão da competição.

Todos aqueles que acompanham o esporte tem ao menos algum interesse na previsão de resultados dessas partidas. Discussões dessas previsões constituem parte essencial da cultura esportiva, desde aquelas pouco confiáveis praticadas entre fãs do esporte, até conversas mais aprofundadas entre especialistas veiculadas ou não em programas esportivos.

Ao longo dos últimas décadas, diversos métodos estatísticos foram propostos para modelar os resultados de partidas de futebol. Em um dos trabalhos pioneiros nesta área, Lee (1997) propõe um modelo para o do campeonato inglês da temporada 1994/1995. Em seu artigo, o autor considera o número de gols de cada time em uma partida como sendo variáveis independentes com distribuição Poisson, sendo que o número médio de gols é explicado por efeitos de ataque e defesa dos times no confronto, além de um efeito associado ao local da partida (em casa/neutro/fora de casa). Dyte e Clarke (2000) propuseram uma modificação desse modelo para seleções nacionais. Ao invés de considerar coeficientes de ataque e defesa, utilizaram o ranking FIFA de cada um dos times. De forma similar, Suzuki *et al.* (2010) consideraram também o número médio de gols dos times dependendo das pontuações FIFA das equipes, mas adotaram uma abordagem bayesiana

em que a distribuição a priori do número médio de gols é obtida a partir de opiniões de experts sobre os placares dos jogos usando o método chamado de *power prior*. Seguindo uma linha mais voltada ao aprendizado de maquina, Groll *et al.* (2019) criou um modelo "híbrido" a partir da utilização dos resultados de um primeiro modelo em um segundo modelo. O primeiro modelo encontra parâmetros que representam a habilidade dos times, sendo preditos por métodos de ranqueamento dessas habilidades a partir da distribuição Poisson. Esses parâmetros por sua vez servem como uma das covariáveis de uma Floresta Aleatória, que é utilizada de fato para previsões das probabilidades de vitória, empate e derrota em uma partida. Todos os modelos obtiveram performances satisfatórias para os campeonatos que tentaram prever.

O objetivo deste trabalho é construir um modelo que fornece previsões probabilísticas para os jogos da Copa do Mundo de 2022 e que seja capaz de acertar os resultados dos jogos. Particularmente, nos inspiramos nos modelos de Lee (1997) e Suzuki *et al.* (2010). Como no caso do primeiro autor, consideramos a distribuição do número de gols como variáveis aleatórias Poisson parametrizadas por coeficientes de ataque e defesa de cada time. A modificação virá do fato de utilizarmos as pontuações dos jogadores no jogo FIFA 23 para o cálculo desses coeficientes. Assim como o segundo artigo, o procedimento bayesiano será aplicado para encontrarmos a distribuição marginal dos dados, porém ao invés de uma *power prior* com opiniões de especialistas, iremos considerar uma priori gamma que considera as probabilidades de derrota, empate e vitória dos times.

Capítulo 2

Material e Métodos

O modelo baseia-se em três fontes de informação: pontuações de ataque e defesa dos jogadores obtidas no jogo FIFA 23, *odds* de sites de aposta esportiva e número de gols marcados pelo time durante a competição.

Dado a disponibilidade (ou não) dessas informações, temos dois modelos: Sem Jogos (subseção 2.1.1) e Com Jogos (subseção 2.1.2). O primeiro indica o modelo mais simples, utilizado para prever os jogos da primeira rodada da Fase de Grupos. Sua maior simplicidade dá-se pelo fato de que na previsão desses jogos ainda não temos nenhuma informação de confrontos passados, seja do número de gols, seja das pontuações de ataque e defesa dos adversários nesses jogos. O segundo modelo, por outro lado, considera esse histórico das partidas que já ocorreram, sendo possível utilizá-lo para prever o torneio após ocorrerem os jogos da primeira rodada da Fase de Grupos.

Com ambos os modelos definidos, estabelecemos a estrutura utilizada para as previsões: antes de cada rodada, as informações disponíveis até o momento são utilizadas para prever o restante da competição. O Modelo Sem Jogos é utilizado para simular os jogos da primeira rodada da Fase de Grupos até a final, com 50000 iterações. Após esses jogos ocorrerem, antes da segunda rodada o Modelo com Jogos é aplicado para simular todo o torneio a partir da segunda rodada, novamente com 50000 iterações. O Modelo Com Jogos é usado antes de cada uma das rodadas até a final e disputa de terceiro lugar, simulando sempre todo o restante do torneio.

As *odds* dos sites de aposta esportiva são incorporadas ao modelo a partir da estimação dos parâmetros da distribuição a priori de cada modelo, com seu processo de estimação explicada na subseção 2.1.3. A cada rodada as *odds* são atualizadas pelos sites de aposta, e será sempre utilizada a versão mais recente delas para estimar os parâmetros

da distribuição a priori do modelo utilizado para a simulação em questão.

A seção 2.2 trata da construção dos escores de ataque e defesa dos times a partir dos escores de ataque e defesa dos jogadores e como eles são considerados no modelo.

Este capítulo encerra-se com uma definição mais detalhada das *odds* e seu processo de conversão em probabilidades (seções 2.3 e 2.4) e a definição dos escores utilizados para avaliação da performance do modelo em cada rodada (seção 2.5).

2.1 Modelo Preditivo Bayesiano

2.1.1 Sem jogos (Modelo "A Priori") .

Verossimilhança

Considere o placar de uma partida entre os times A e B seja dado pelo vetor aleatório $(X_{A,B}, X_{B,A})$, em que $X_{A,B}$ é o número de gols que o time A faz em B e $X_{B,A}$ é o número de gols que B faz em A. Vamos supor que as variáveis aleatórias $X_{A,B}$ e $X_{B,A}$ tem distribuição de Poisson independentes condicionalmente ao conhecimento dos seus respectivos valores médios $\mu_{A,B}$ e $\mu_{B,A}$, isto é,

$$P\left(X_{A,B} = x_A \,|\, \mu_{A,B}\right) = e^{-\mu_{A,B}} \frac{\mu_{A,B}^{x_A}}{x_A!}, \qquad x_A = 0, 1, \ldots, \tag{2.1}$$

$$P\left(X_{B,A} = x_B \,|\, \mu_{B,A}\right) = e^{-\mu_{B,A}} \frac{\mu_{B,A}^{x_B}}{x_B!}, \qquad x_B = 0, 1, \ldots. \tag{2.2}$$

Os valores médios $\mu_{A,B}$ e $\mu_{B,A}$ são modelados como

$$\mu_{A,B} = \lambda_A \frac{ATT_A}{DEF_B},$$
$$\mu_{B,A} = \lambda_B \frac{ATT_B}{DEF_A}.$$

em que ATT_A (ATT_B) representa a pontuação de ataque do time A (B), DEF_B (DEF_A) a pontuação de defesa do time B (A), λ_A (λ_B) o número médio de gols que A (B) faz contra um time B (A) tal que $DEF_B = ATT_A$ ($DEF_A = ATT_B$). A razão $\frac{ATT_A}{DEF_B}$ determina um fator multiplicativo para a taxa de gols λ_A do time A, isto é, quanto maior for ATT_A maior é o número médio $\mu_{A,B}$ e quanto maior DEF_B menor é o número

médio $\mu_{A,B}$. Uma interpretação similar é válida para $\mu_{B,A} = \lambda_B \frac{ATT_B}{DEF_A}$.

Distribuição a priori

Assume-se que $\lambda_A \sim \text{Gamma}(\alpha_A, \beta_A)$ e $\lambda_B \sim \text{Gamma}(\alpha_B, \beta_B)$. Mais especificamente, temos

$$f(\lambda_A) = \frac{\beta_A^{\alpha_A}}{\Gamma(\alpha_A)} \lambda_A^{\alpha_A - 1} e^{-\beta_A \lambda_A}, \qquad \alpha_A, \beta_A > 0, \tag{2.3}$$

com expressão equivalente para λ_B, dada por

$$f(\lambda_B) = \frac{\beta_B^{\alpha_B}}{\Gamma(\alpha_B)} \lambda_B^{\alpha_B - 1} e^{-\beta_B \lambda_B}, \qquad \alpha_B, \beta_B > 0. \tag{2.4}$$

Temos que $\alpha_A, \beta_A, \alpha_B, \beta_B$ contêm as informações das *odds*. Para estimá-los, convertem-se as todas as *odds* em probabilidades, como descrito na seção 2.4. Suponha-se então uma partida entre os times A e B. Fixam-se as probabilidades de vitória dos times como sendo os valores das densidades $f(\lambda_A)$ e $f(\lambda_B)$, e o método L-BFGS, descrito na subseção 2.1.3, encontra os valores de $\alpha_A, \beta_A, \alpha_B, \beta_B$ associados a estas probabilidades.

Posteriori

Como $X_A|\lambda_A \sim \text{Poisson}(\lambda_A \frac{ATT_A}{DEF_B})$ e $\lambda_A \sim \text{Gamma}(\alpha_A, \beta_A)$, é possível demonstrar que $\lambda_A|X_A = x_A \sim \text{Gamma}(\alpha_A + x_A, \frac{ATT_A}{DEF_B} + \beta_A)$, ou seja,

$$f(\lambda_A|X_A = x_A) = \frac{(\beta_A + \frac{ATT_A}{DEF_B})^{\alpha_A + x_A}}{\Gamma(\alpha_A + x_A)} \lambda_A^{\alpha_A + x_A - 1} e^{-\left(\beta_A + \frac{ATT_A}{DEF_B}\right)\lambda_A}. \tag{2.5}$$

Aplicando o mesmo processo para o time B, temos

$$f(\lambda_B|X_B = x_B) = \frac{(\beta_B + \frac{ATT_B}{DEF_A})^{\alpha_B + x_B}}{\Gamma(\alpha_B + x_B)} \lambda_B^{\alpha_B + x_B - 1} e^{-\left(\beta_B + \frac{ATT_B}{DEF_A}\right)\lambda_B}. \tag{2.6}$$

A demonstração feita para o time A encontra-se no Apêndice A.1.

Preditiva a priori

Com as distribuições de $\lambda_A|X_A = x_A$ e $X_A|\lambda_A$, é possível encontrar a distribuição preditiva a priori do número de gols, dada por $X_A \sim \text{BN}\left(\alpha_A, \frac{\beta_A}{\beta_A + \frac{ATT_A}{DEF_B}}\right)$, ou seja,

$$P(X_A = x_A) = \begin{cases} \frac{\Gamma(\alpha_A + x_A)}{x_A!\ \Gamma(\alpha_A)} \left(\frac{\frac{ATT_A}{DEF_B}}{\frac{ATT_A}{DEF_B} + \beta_A} \right)^{x_A} \left(\frac{\beta_A}{\frac{ATT_A}{DEF_B} + \beta_A} \right)^{\alpha_A} & , \text{se } x_A = 0, 1, \ldots \\ 0 & , \text{caso contrário.} \end{cases} \tag{2.7}$$

Para o time B, temos

$$P(X_B = x_B) = \begin{cases} \frac{\Gamma(\alpha_B + x_B)}{x_B!\ \Gamma(\alpha_B)} \left(\frac{\frac{ATT_B}{DEF_A}}{\frac{ATT_B}{DEF_A} + \beta_B} \right)^{x_B} \left(\frac{\beta_A}{\frac{ATT_B}{DEF_A} + \beta_B} \right)^{\alpha_B} & , \text{se } x_B = 0, 1, \ldots \\ 0 & , \text{caso contrário.} \end{cases} \tag{2.8}$$

A demonstração para o time A encontra-se no Apêndice A.2.

2.1.2 Com Jogos (Modelo "A Posteriori")

Distribuição a priori

A distribuição a priori utilizada é a mesma do modelo sem jogos, dada pela expressão 2.3. A cada rodada os sites de aposta atualizam suas *odds* para os confrontos que irão ocorrer, e estas *odds* são utilizadas para estimar os parâmetros $\alpha_A, \beta_A, \alpha_B, \beta_B$ para todos os confrontos, como já mencionado no modelo sem jogos.

Verossimilhança

Suponha-se que até o momento, o time A jogou n partidas. Na primeira, marcou X_{A1} gols, na segunda, marcou X_{A2}, e assim por diante até a partida X_{An}. Sendo assim, a verossimilhança que incorpora esses números de gols ao modelo é dada por

$$P(X_{A1}, X_{A2}, \ldots, X_{An} | \lambda_A) = \frac{e^{(-\sum_{i=1}^{n} \mu_{Ai})} \mu_{Ai}^{\sum_{i=1}^{n} x_{Ai}}}{\prod_{i=1}^{n} x_{Ai}!}. \tag{2.9}$$

Fazendo o mesmo para o time B, temos

$$P(X_{B1}, X_{B2}, \ldots, X_{Bn} | \lambda_B) = \frac{e^{(-\sum_{i=1}^{n} \mu_{Bi})} \mu_{Bi}^{\sum_{i=1}^{n} x_{Bi}}}{\prod_{i=1}^{n} x_{Bi}!}. \tag{2.10}$$

Posteriori

Temos que $\lambda_A|X_{A1},\ldots,X_{An} \sim Gama\left(\sum_{i=1}^{n} x_{Ai}+\alpha_A, \sum_{i=1}^{n}\frac{ATT_A}{DEF_{Bi}}+\beta_A\right)$, ou seja,

$$P(\lambda_A|X_{A1},\ldots,X_{An}) = \frac{\left(\sum_{i=1}^{n}\frac{ATT_A}{DEF_{Bi}}+\beta_A\right)^{\sum_{i=1}^{n}x_{Ai}+\alpha_A}}{\Gamma(\sum_{i=1}^{n}x_{Ai}+\alpha_A)}\lambda_A^{\sum_{i=1}^{n}x_{Ai}+\alpha_A-1}e^{-\left(\sum_{i=1}^{n}\frac{ATT_A}{DEF_{Bi}}+\beta_A\right)\lambda_A}, \tag{2.11}$$

que é obtida a partir da priori e da verossimilhança, e sua demonstração encontra-se no Apêndice A.3.

Preditiva a Posteriori

Considerando $\mathbf{X} = [X_{A1},\ldots,X_{An}]$ a matriz linha com o número de gols do time A marcados em suas partidas anteriores e X_A o número de gols do próximo jogo, temos

$$\begin{aligned} P(X_A|\mathbf{X}=\mathbf{x}) = &\left(\frac{\sum_{i=1}^{n}\frac{ATT_A}{DEF_{Bi}}+\beta_A}{\sum_{i=1}^{n}\frac{ATT_A}{DEF_{Bi}}+\frac{ATT_A}{DEF_B}+\beta_A}\right)^{\sum_{i=1}^{n}x_{Ai}+\alpha_A} \\ &\times\left(\frac{1}{\sum_{i=1}^{n}\frac{ATT_A}{DEF_{Bi}}+\frac{ATT_A}{DEF_B}+\beta_A}\right)^{x_a}\binom{\sum_{i=1}^{n}x_{Ai}+\alpha_A+x_a-1}{x} \rightarrow \\ \rightarrow \therefore &\boxed{X_A|\mathbf{X}=\mathbf{x} \sim BN\left(\sum_{i=1}^{n}x_{Ai}+\alpha_A, \frac{\sum_{i=1}^{n}\frac{ATT_A}{DEF_{Bi}}+\beta_A}{\sum_{i=1}^{n}\frac{ATT_A}{DEF_{Bi}}+\frac{ATT_A}{DEF_B}+\beta_A}\right)}. \end{aligned} \tag{2.12}$$

Sua demonstração pode ser verificada no Apêndice A.4.

2.1.3 Encontrando os parâmetros $\alpha_A, \beta_A, \alpha_B, \beta_B$ a Distribuição a priori

Seja $R = \{A, E, B\}$ o conjunto dos possíveis resultados de uma partida, em que A representa a vitória do time A, B a vitória do time B e E o empate. Considere então $P = (P_A, P_E, P_B)$ o vetor de probabilidades preditas associadas ao conjunto R. Considere ainda $P^{(o)} = (P_A^{(o)}, P_E^{(o)}, P_B^{(o)})$ como o vetor de probabilidades obtido pelas *odds*. Para encontrar uma estimativa inicial dos parâmetros $(\alpha_A, \beta_A, \alpha_B, \beta_B)$, minimiza-se a função

$$(\alpha_A^*, \beta_A^*, \alpha_B^*, \beta_B^*) = \underset{(\alpha_A,\beta_A,\alpha_B,\beta_B)}{\operatorname{argmin}} \sum_{i \in R} (P_i(\alpha_A, \beta_A, \alpha_B, \beta_B) - P_i^{(o)})^2, \tag{2.13}$$

ou seja, o vetor de estimativas $(\alpha_A^*, \beta_A^*, \alpha_B^*, \beta_B^*)$ é dado pelos valores de $(\alpha_A, \beta_A, \alpha_B, \beta_B)$ cujas probabilidades preditas usando a distribuição preditiva a priori $P_i(\alpha_A, \beta_A, \alpha_B, \beta_B)$ mais se aproximam do vetor de probabilidades $P^{(o)} = (P_A^{(o)}, P_E^{(o)}, P_D^{(o)})$ proveniente das *odds*. Para essa otimização, será utilizado o método L-BFGS (*Limited Memory* BFGS). O método, proposto por Liu e Nocedal (1989), consiste em uma modificação do BFGS (*Broyden-Fletcher-Goldfarb-Shanno*), proposto por Fletcher (1987). Ambos os métodos são descritos a seguir, com ênfase no L-BFGS.

L-BFGS

O método L-BFGS proposto por Liu e Nocedal (1989) pertence à família de métodos de otimização quase-Newtonianos. Dada uma função $f(\mathbf{x})$, com $\mathbf{x} \in \mathbb{R}^k, k \in \mathbb{N}$, é necessário apenas o gradiente $\nabla f(\mathbf{x})$ para minimizarmos (ou maximizarmos) $f(\mathbf{x})$, ou seja, apenas o vetor com as primeiras derivadas parciais de $f(\mathbf{x})$. Isso difere dos métodos Newtonianos, que exigem o Hessiano $\mathbf{H}(\mathbf{x})$, ou seja, a matriz que contém as derivadas parciais de cada primeira derivada parcial obtida em $\nabla f(\mathbf{x})$. Os métodos quase-Newtonianos portanto utilizam uma aproximação de $\mathbf{H}(\mathbf{x})$ obtida a partir de $\nabla f(\mathbf{x})$. Resumidamente, o método utiliza uma aproximação de $\mathbf{H}(\mathbf{x})$ menor que uma matriz n x n (sendo n o número de variáveis no vetor $\mathbf{x}$), dado que para valores altos de n a minimização pode tornar-se intratável computacionalmente.

Dada a estimativa de $\mathbf{H}(\mathbf{x})$, denominada $\mathbf{Q}(\mathbf{x})$, o método busca atualizar os valores em $\mathbf{x}$ de forma iterativa até que tenhamos uma aproximação considerada suficiente do mínimo de $f(\mathbf{x})$. A atualização de $\mathbf{x}$ do passo k ao passo $k+1$ ocorre da seguinte forma:

$$\begin{aligned} \mathbf{x}^{(k+1)} &= \mathbf{x}^{(k)} - \mathbf{Q}^{(k)} \nabla f(\mathbf{x}) = \\ &= \mathbf{x}^{(k)} - \mathbf{d}^{(k)}, \end{aligned}$$

em que $\mathbf{d}^{(k)} = \mathbf{Q}^{(k)} \nabla f(\mathbf{x})$ é a direção de variação. Iremos iterar por k passos até que o mínimo seja suficientemente aproximado. Para facilitar a representação das contas, iremos considerar $\mathbf{g} = \nabla f(\mathbf{x})$ e

$$\boldsymbol{\gamma}^{(k+1)} = \mathbf{g}^{(k+1)} - \mathbf{g}^{(k)}$$
$$\boldsymbol{\delta}^{(k+1)} = \mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}$$

O método irá encontrar primeiramente $\mathbf{W}$, uma matriz n x m em que cada uma das m colunas é encontrada de forma iterativa. Na primeira iteração, encontra-se a m-ésima coluna de $\mathbf{W}$, dada por $\mathbf{w}^{(m)} = g$. Os outros vetores, calculados de $m-1$ até 1, são dados por

$$\mathbf{w}^{(i)} = \mathbf{w}^{(i+1)} - \frac{\left(\boldsymbol{\delta}^{(i+1)}\right)^T \mathbf{w}^{(i+1)}}{\left(\boldsymbol{\gamma}^{(i+1)}\right)^T \boldsymbol{\delta}^{(i+1)}} \boldsymbol{\gamma}^{(i+1)}$$

.

O valor de m pode ser definido pelo usuário, e indica que os os últimos m valores obtidos de $\boldsymbol{\gamma}$ e $\boldsymbol{\delta}$ serão usados na estimativa. Nesta notação, m indica valor mais recente e 1 o mais antigo. Na prática, o valor de m utilizado é o valor ótimo definido pela função *optim* do R que implementa o L-BFGS, escolhido automaticamente de forma que a função $f(\mathbf{x})$ seja minimizada com sucesso.

Cada vetor $w^{(i)}$ por sua vez será utilizado para estimar vetores $z^{(i)}$, $i = 1$ até m pela fórmula

$$\mathbf{z}^{(i)} = \mathbf{z}^{(i-1)} + \boldsymbol{\delta}^{(i-1)} \left(\frac{\left(\boldsymbol{\delta}^{(i-1)}\right)^T \mathbf{w}^{(i-1)}}{\left(\boldsymbol{\gamma}^{(i-1)}\right)^T \boldsymbol{\delta}^{(i-1)}} - \frac{\left(\boldsymbol{\gamma}^{(i-1)}\right)^T \mathbf{z}^{(i-1)}}{\left(\boldsymbol{\gamma}^{(i-1)}\right)^T \boldsymbol{\delta}^{(i-1)}} \right)$$

,

que resulta na direção $\mathbf{d} = -\mathbf{z}^{(m)}$.

Nota-se que para $\mathbf{z}^{(1)}$, seriam necessários $\boldsymbol{\delta}^{(0)}$ e $\boldsymbol{\gamma}^{(0)}$, o que não temos. Para o primeiro passo, é assumido que

$$\mathbf{Q}^{(1)} = \frac{\boldsymbol{\gamma}^{(1)} \left(\boldsymbol{\delta}^{(1)}\right)^T}{\left(\boldsymbol{\gamma}^{(1)}\right)^T \boldsymbol{\gamma}^{(1)}}$$

e assim $\mathbf{z}^{(1)} = \mathbf{Q}^{(1)} \mathbf{z}^{(1)}$.

2.2 Construção das Pontuações de Ataque e Defesa

2.2.1 Pontuações de Ataque e Defesa do Jogo Eletrônico FIFA 23

No jogo eletrônico FIFA, a força de cada time é dada pelo conjunto de atributos de cada jogador do elenco. Esses atributos são bastante específicos e em grande quantidade. Por exemplo, temos aqueles que medem habilidades específicas ao futebol tais como passes curtos, passes longos, cobranças de falta, dribles, entre outros. Temos ainda atributos de condicionamento físico, como agilidade, equilíbrio, tempo de reação e resistência. Para construção das pontuações de ataque e defesa de cada equipe, será considerada neste trabalho a pontuação geral de cada atleta, que pode ser entendida como um número que tenta representar essas habilidades de forma conjunta. Uma extensão natural deste trabalho, portanto, seria construir as pontuações de ataque e defesa considerando os atributos individualmente. No caso de um jogador da Copa do Mundo, essa pontuação geral é um inteiro no intervalo $[56, 91]$. O jogador com menor pontuação presente na Copa tem uma pontuação de valor 56. Neste caso, é o atleta Asiri Haitham, da Arábia Saudita (KSA). Da mesma forma, os jogadores com uma pontuação de valor 91 são Kylian Mbappé (FRA), Kevin De Bruyne (BEL), Robert Lewandowski (POL) e Lionel Messi (ARG). A distribuição das pontuações é representada no histograma abaixo:

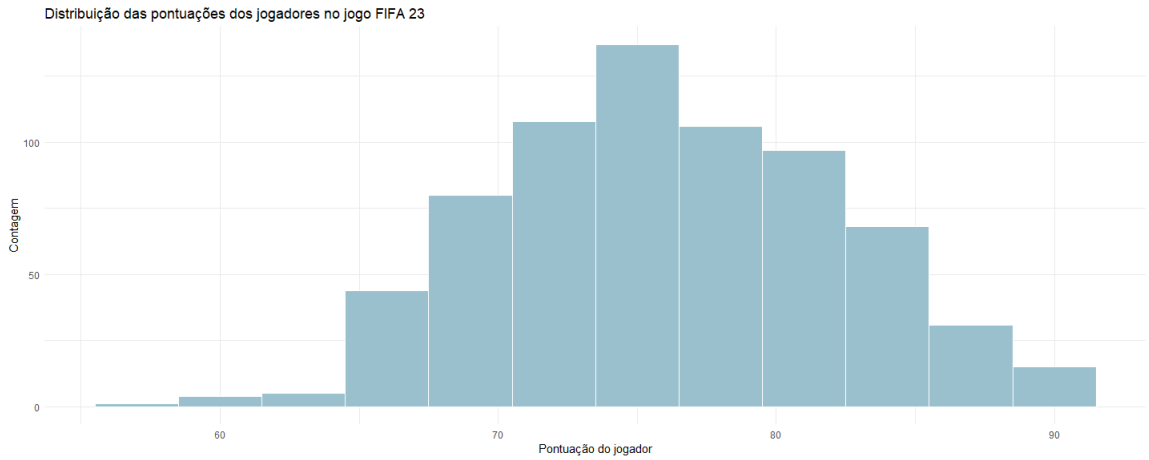

Figura 2.1: Distribuição da pontuação geral dos jogadores (*Overall*) no jogo Fifa 23.

Para calcular as pontuações de ataque (ATT) e defesa (DEF) de uma equipe serão utilizadas as pontuações de cada um dos 26 jogadores convocados de cada seleção. Cada jogador é classificado inicialmente em uma de três categorias: ataque, defesa ou meio-campo. Em seguida é feita a média das pontuações dentro de cada categoria para cada

time. O gráfico a seguir ilustra essa divisão para cada seleção:

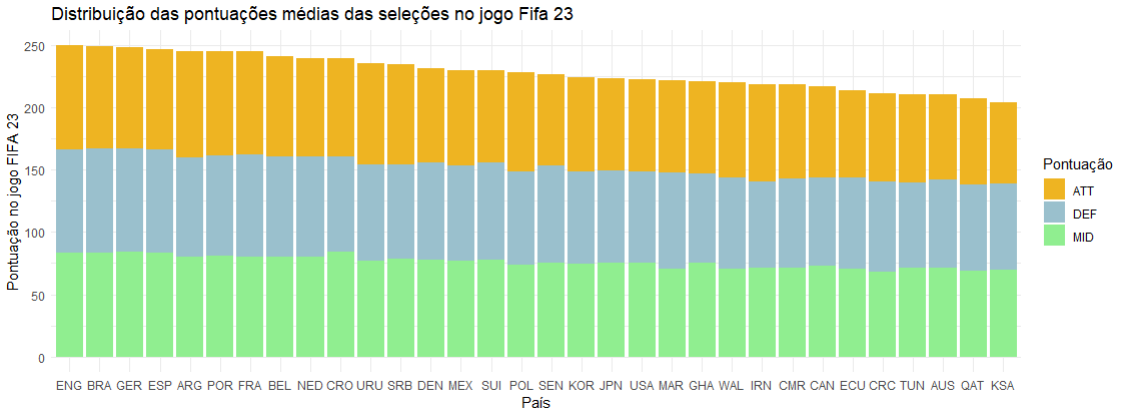

Figura 2.2: Pontuações de ataque, defesa e de meio-campo de cada seleção após divisão inicial dos jogadores entre estes setores.

A partir da Figura 2.2, que ordena os times de acordo com a maior soma das pontuações de ataque, defesa e meio-campo, nota-se que o jogo FIFA 23 atribui maior quantidade de pontos aos jogadores das seleções consideradas mais fortes (Inglaterra, Brasil, Alemanha, Espanha, Argentina), assim como atribui menos pontos às seleções mais fracas (Arábia Saudita, Catar, Austrália, Tunísia e Costa Rica).

Feita essa divisão inicial, é necessário dividir as pontuações de meio-campo entre as categorias de ataque e defesa. Essa divisão busca adaptar-se ao modelo de Suzuki *et al.* (2010), que considera apenas pontuações de ataque (ATT) e defesa (DEF). Para fazê-la, somam-se as pontuações de meio-campo com as pontuações de ataque e depois divide-se esse valor por dois, resultando na nova pontuação de ataque. O mesmo é feito com as pontuações de defesa. Essa divisão é representada na figura a seguir:

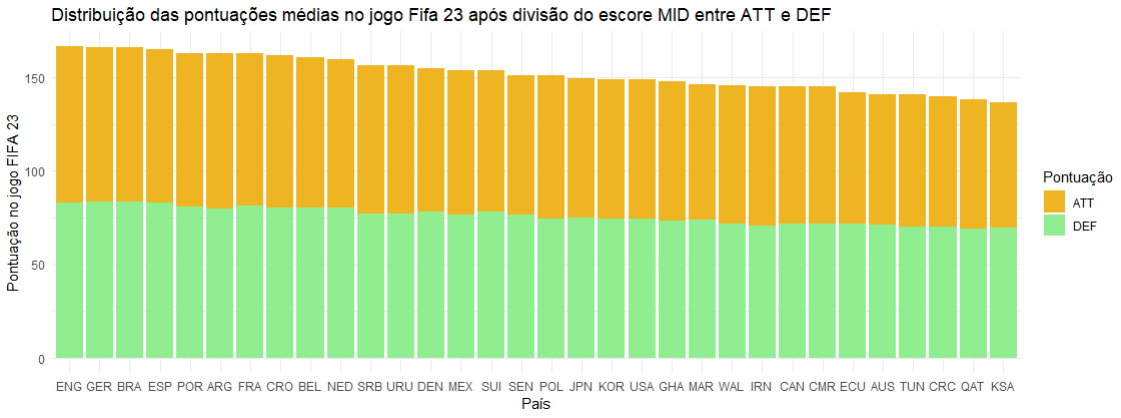

Figura 2.3: Pontuações de ataque e defesa de cada seleção após divisão da pontuação de meio-campo entre os outros dois.

Os intervalos obtidos após a divisão são $ATT \in [67.31, 83.46]$ e $DEF \in [68.93, 83.74]$. Ao realizar uma simulação do torneio utilizando esses valores, percebeu-se que jogos com um número aberrante de gols eram produzidos. Por exemplo, haviam partidas em que um time fazia 20 gols, sendo que o número máximo de gols feitos por um time nas duas edições anteriores da copa (2018 e 2014) foi de 7 gols. Sendo assim, buscou-se controlar esse impacto por meio da padronização das pontuações de ATT e DEF, explicado na subseção a seguir. Outro motivo para a realização da padronização é a produção de pontuações de fácil interpretação. Maiores detalhes sobre esses motivos assim como a explicação da padronização utilizada encontram-se na subseção a seguir.

Nota-se que uma extensão natural deste trabalho é estudar a divisão das pontuações dos jogadores de meio-campo considerando a característica individual deste setor em cada equipe. Por exemplo, se o meio-campo de um time tem característica mais ofensiva, pode ser feita uma divisão de forma que mais pontos desses jogadores sejam alocados para seu escore de ataque. O mesmo raciocínio vale caso seu meio-campo seja mais defensivo, ou seja, alocaria-se maior parcela dos pontos desses jogadores à pontuação de defesa do time.

2.2.2 Padronização das pontuações de Ataque e Defesa

Para determinação de um intervalo de variação para as pontuações, vamos considerar o impacto que a pontuação tem no número médio de gols $\mu_{A,B}$ que uma equipe A faz em uma equipe B. Lembremos que

$$\mu_{A,B} = \lambda_A \frac{ATT_A}{DEF_B},$$

em que ATT_A é a pontuação de ataque da equipe A e DEF_B é a pontuação de defesa de B. Observe que as pontuações de ataque e defesa afetam o número médio usual de gols que A faz em uma equipe com qualidade similar (λ_A) por meio de um fator multiplicativo $\gamma_{A,B} = ATT_A/DEF_B$. Observe que

$$\begin{cases} \mu_{A,B} > \lambda_A, & \text{se } \gamma_{A,B} > 1, \\ \mu_{A,B} = \lambda_A, & \text{se } \gamma_{A,B} = 1, \\ \mu_{A,B} < \lambda_A, & \text{se } \gamma_{A,B} < 1, \end{cases}$$

isto é, quando $\gamma_{A,B} = 1$ o número médio $\mu_{A,B}$ permanece igual a λ_A, quando $\gamma_{A,B} > 1$ o número médio $\mu_{A,B}$ aumenta e quando $\gamma_{A,B} < 1$ o número médio $\mu_{A,B}$ diminui.

A partir desta análise, percebemos que o intervalo de variação das pontuações das equipes tem um papel importante na modelagem do número médio de gols. Propomos encontrar um intervalo para as pontuações da forma $[1-\delta, 1+\delta]$, $0 < \delta < 1$. A escolha de um intervalo que seja simétrico em torno de 1 foi feita para permitir algumas interpretações interessantes para as pontuações:

(i) uma equipe com escore de ataque (defesa) igual a 1 representa uma equipe com qualidade de ataque (defesa) mediana;

(ii) quanto maior a pontuação de ataque (defesa) maior o poder ofensivo (defensivo) da equipe;

(iii) quanto menor o valor da pontuação de ataque (defesa) menor o poder ofensivo (defensivo) da equipe.

Considerando ATT_i (DEF_i) a pontuação de ataque (defesa) da equipe i, $i = 1, \ldots, T$, devemos ter

$$\begin{aligned} 1-\delta < ATT_i < 1+\delta, &\qquad \text{para todo } i = 1, \ldots, T, \\ 1-\delta < DEF_i < 1+\delta, &\qquad \text{para todo } i = 1, \ldots, T, \end{aligned}$$

para um valor $\delta \in [0, 1]$ a ser determinado.

Além da interpretabilidade das pontuações, este intervalo permite controlar o impacto das pontuações na modelagem, pois o valor máximo possível de $\gamma_{A,B} = \frac{ATT_A}{DEF_B}$ ocorre quando $ATT_A = 1 + \delta$ e $DEF_B = 1 - \delta$. Note que quanto maior o valor de δ, maior o valor de $\gamma_{A,B}$ e portanto há um maior "peso" dado às pontuações de ATT e DEF na modelagem. Nota-se portanto que a escolha de δ é crucial para determinar a qualidade do modelo. Essa razão máxima possível será denominada k de agora em diante, ou seja,

$$k = \frac{1+\delta}{1-\delta}. \tag{2.14}$$

O modo como δ é encontrado está explicitado mais adiante, mas por enquanto vale ressaltar que ele será encontrado a partir de k. Sendo assim, é necessário isolar δ em $k = \frac{1+\delta}{1-\delta}$, obtendo a seguinte expressão:

$$\delta = \frac{k-1}{k+1}. \tag{2.15}$$

Sendo assim, o novo intervalo das pontuações, agora em função de k, é dado por

$$\begin{aligned}[1-\delta, 1+\delta] &= \left[1 - \frac{k-1}{k+1}, 1 + \frac{k-1}{k+1}\right] = \\ &= \left[\frac{2}{k+1}, \frac{2k}{k+1}\right], \qquad \text{para algum } k \geq 1. \end{aligned} \tag{2.16}$$

Dado os limites inferior e superior das pontuações, é possível padronizar cada escore observado E de acordo com o procedimento

$$E' = \frac{2}{k+1} + \frac{2(k-1)}{k+1} \times \frac{E - E_{min}}{E_{max} - E_{min}}, \tag{2.17}$$

em que E é a pontuação original, E_{max} é o valor máximo observado das pontuações considerando ATT e DEF, E_{min} é o valor mínimo e E' é a pontuação padronizada.

Encontrando o Valor de δ

Para encontrar k^*, o valor de k de fato utilizado no modelo, realizam-se simulações usando vários valores de k. Como definido na expressão (2.14), quanto maior o valor de k, maior será a razão $\frac{1+\delta}{1-\delta}$. Isso vale também para as razões entre os outras pontuações dos times, que são valores dentro desse intervalo, ou seja, suas razões também aumentam. Pode-se dizer então que aumentar o valor de k aumenta a distância entre as pontuações dos times, e portanto maior o peso desses escores no modelo. Sendo assim, busca-se um valor de k suficientemente grande para que as pontuações diferenciem os times, mas não tão grande ao ponto de desconsiderar as outras fontes de informação do modelo (*odds* dos sites de aposta e número de gols marcados ao longo da Copa de 2022).

Para cada k, é feita a conversão das pontuações de ATT e DEF de todos os times segundo a expressão (2.17) e depois é realizado o confronto entre o time com maior escore de ataque e o time com pior escore de defesa. O melhor escore de ataque pertence à Inglaterra (ENG). Na escala original, ou seja, antes de usar a transformação para o intervalo $[1-\delta, 1+\delta]$, tem-se que $ATT_{ENG} = 83.46$, enquanto o time com pior escore de defesa é o Catar (QAT), com valor $DEF_{QAT} = 68.93$. Foi considerado que, mesmo neste confronto em que ocorre a maior disparidade entre ataque e defesa, a probabilidade da Inglaterra

marcar um grande número de gols sobre o Catar é pequena. Mais especificamente, consideramos que a probabilidade da Inglaterra marcar mais de 5 gols no Catar é de 0.05. Este valor apoia-se no que foi observado nas duas últimas edições da Copa do Mundo: na edição de 2018, das 64 partidas disputadas, apenas em 3 delas (ou 5%, aproximadamente) houve um número maior que 5 gols feitos por um único time: Rússia x Arábia Saudita (5 x 0), Bélgica x Tunísia (5 x 2) e Inglaterra x Panamá (6 x 1). O Mesmo ocorreu na Copa de 2014, que além do jogo do Brasil na semifinal, apenas em mais dois jogos um número maior que 4 gols ocorreu: Holanda x Espanha (5 x 1) e França x Suíça (5 x 2).

Portanto, o valor ótimo considerado para k, k^*, é aquele que produz uma probabilidade de 0.05 da Inglaterra marcar mais que 5 gols no Catar. Matematicamente, temos que encontrar k tal que $P(X_{ENG,QAT} \geq 5) \leq 0.05$. Desenvolvendo a expressão de acordo com a verossimilhança do modelo em 2.1.1, temos que encontrar k tal que

$$\begin{aligned} P(X_{ENG,QAT} \geq 5 | \mu_{ENG,QAT,k}) \leq 0.05 = \\ 1 - P(X_{ENG,QAT} < 5 | \mu_{ENG,CAT,k}) \leq 0.05 = \\ 1 - \sum_{x=0}^{4} e^{-\mu_{ENG,QAT}} \frac{\mu_{ENG,QAT,k}^{x}}{x!} \leq 0.05 \end{aligned}$$

em que $\mu_{ENG,CAT,k} = 1.7 \frac{ATT_{ENG,k}}{DEF_{QAT,k}}$ é a média de gols da Inglaterra sobre o Catar sujeita ao valor k, $ATT_{ENG,k}$ é a pontuação de ataque da Inglaterra sujeita ao valor k, $DEF_{QAT,k}$ é a pontuação de defesa do Catar sujeito ao valor k e 1.7 é a média de gols da Inglaterra na edição de 2018 da Copa do Mundo.

Testando com diferentes valores de k, foi obtido o gráfico a seguir:

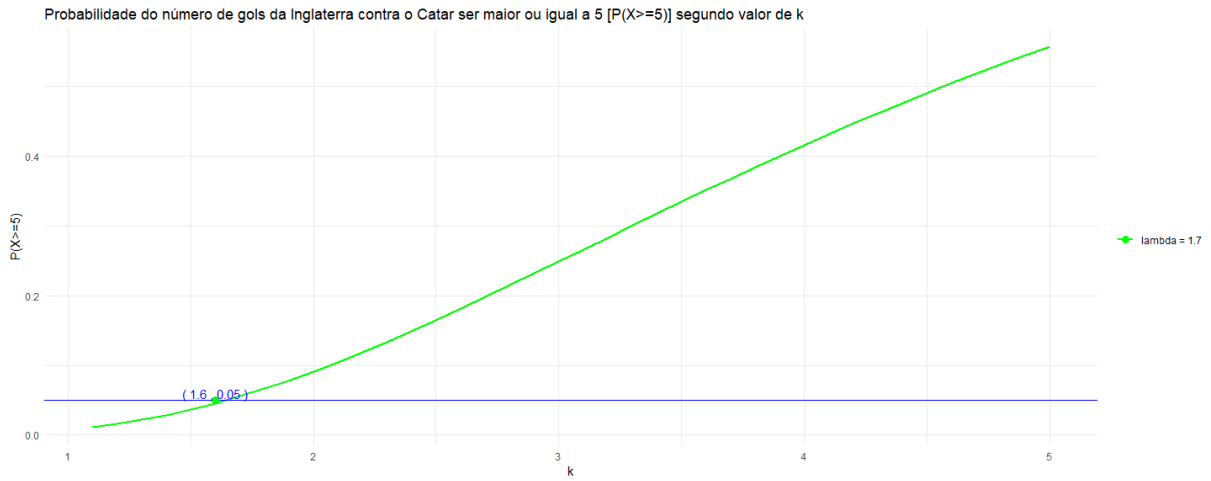

Figura 2.4: A curva em verde ilustra o comportamento da probabilidade do número de gols da Inglaterra contra o Catar ser maior ou igual a 5, de acordo com o valor de k usado para converter as pontuações. A linha azul indica que $P(X_{ENG,QAT} \geq 5) = 0.05$ quando $k = 1.6$.

A partir da Figura 2.4, o valor de k a ser usado no modelo é $k^* = 1.6$. Convertendo todas as pontuações de ATT e DEF segundo k^* e a expressão 2.17, e em seguida confrontando Inglaterra e Catar, temos que

$$\gamma_{ENG,QAT} = 1.70 \times \frac{ATT_{ENG}}{DEF_{QAT}} \tag{2.18}$$

$$= 1.70 \times \frac{1.22}{0.81} = \tag{2.19}$$

$$= 1.70 \times 1.51 \tag{2.20}$$

$$= 2.57. \tag{2.21}$$

Percebe-se que com $k^* = 1.6$, a média de gols da Inglaterra é multiplicada em 1.5 vezes, passando de 1.7 para 2.57 qaundo enfrenta o time do Catar.

Convertendo as pontuações de todos os times segundo $k^* = 1.6$

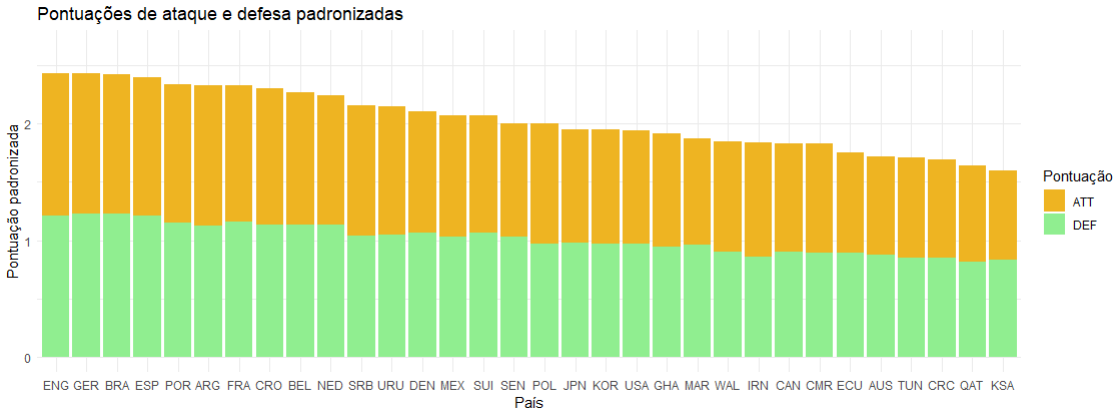

Figura 2.5: Escores de ataque e defesa de cada time após padronização. Estes serão as pontuações utilizados no ajuste do modelo.

O país com maior soma de pontuações de ATT e DEF (ou pontuação média) é a Inglaterra, seguido da Alemanha, Brasil, Espanha e Portugal. O país com a menor soma é a Arábia Saudita, seguida por Catar, Costa Rica, Tunísia e Austrália. Para melhor visualização, foram feitos os gráficos contendo somente um tipo de escore (ATT ou DEF):

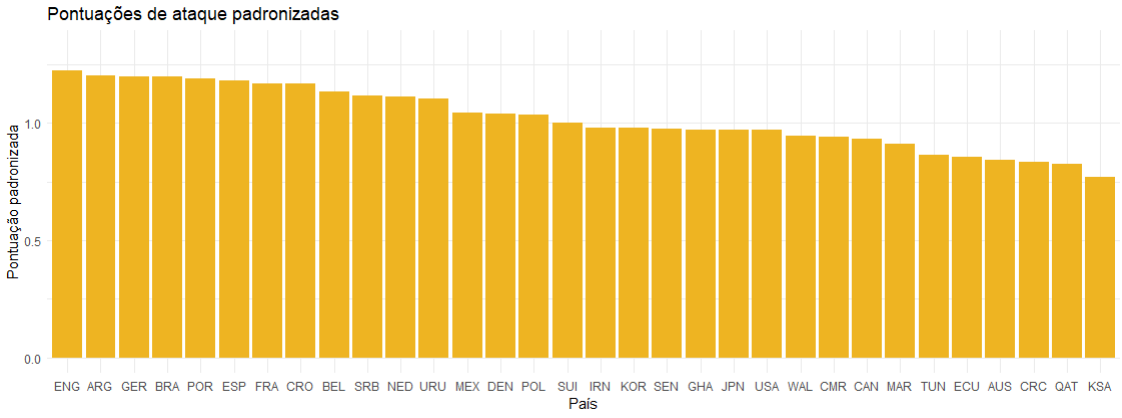

Figura 2.6: Escores de ataque padronizados de cada seleção, que serão utilizados no modelo.

É possível perceber que 5 os melhores escores de ataque são da Inglaterra (1.22), Argentina (1.20), Alemanha (1.20), Brasil (1.20) e Portugal (1.19). Os piores 5 escores pertencem à Arábia Saudita (0.77), Catar (0.83), Costa Rica (0.84), Austrália (0.84) e Equador (0.86). A seguir temos o mesmo gráfico para as pontuações de defesa:

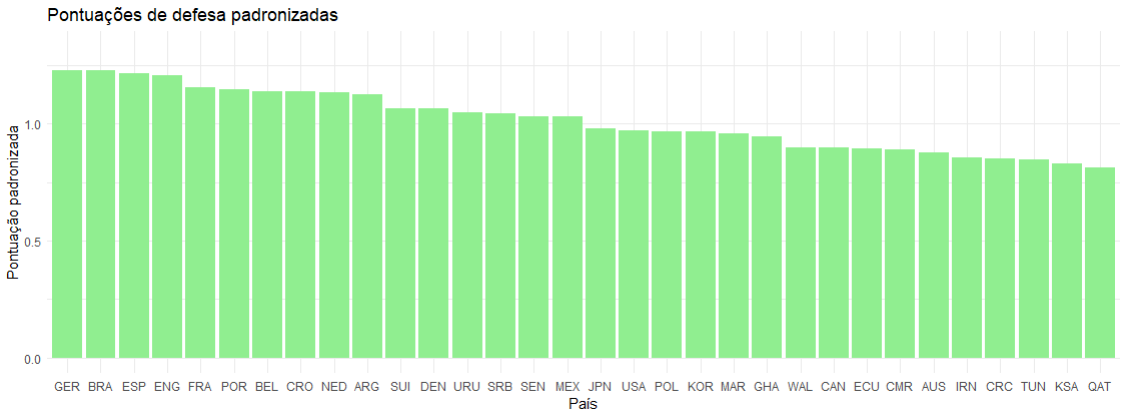

Figura 2.7: Escores de defesa padronizados de cada seleção, que serão utilizados no modelo.

Neste caso as 5 melhores defesas pertencem às seleções da Alemanha (1.23), Brasil (1.23), Espanha (1.21), Inglaterra (1.21) e França (1.16). As 5 piores defesas pertencem ao Catar (0.81), Arábia Saudita (0.83), Tunísia (0.85), Costa Rica (0.85) e Irã (0.86).

Ao analisar as maiores distâncias entre as pontuações de ATT e DEF para o mesmo time, temos o Irã com a maior discrepância ($ATT_{IRN} = 0.98$ e $DEF_{IRN} = 0.86$), seguido por Argentina ($ATT_{ARG} = 1.20$ e $DEF_{ARG} = 1.13$), Sérvia ($ATT_{SRB} = 1.04$ e $DEF_{SRB} = 1.12$), Polônia ($ATT_{POL} = 1.04$ e $DEF_{POL} = 0.96$) e Suíça ($ATT_{SUI} = 1.00$ e $DEF_{SUI} = 1.07$).

2.3 *Odds* de sites de aposta esportiva

Sites de apostas esportivas são plataformas online em que é possível apostar determinado valor monetário para uma variedade cada vez maior de esportes, como futebol, tênis, basquete, beisebol, vôlei, hóquei, entre outros. Para cada esporte, é possível apostar em diversos resultados. No caso do futebol, esses resultados variam do tradicional "time X vai ganhar do time Y" até "time X time fará n faltas (ou n pênaltis, n gols, e assim por diante)". Os sites são geridos por sites de aposta esportiva, empresas que, além das tarefas de gerenciamento do site, têm a responsabilidade de estabelecer os multiplicadores do valor apostado, caso o apostador ganhe. Esses multiplicadores são denominados *odds* ("chances" em português), pois o valor do multiplicador está associado à chance dele ocorrer. Em outras palavras, quanto maior o valor desse multiplicador, menor a chance do evento apostado ocorrer. No caso da Copa do Mundo, um exemplo seria o confronto entre França e Austrália pela primeira rodada do Grupo D, em que os sites de aposta

esportiva provavelmente atribuem um multiplicador maior para a Austrália e outro menor para a França, dado fatores como desempenho do país em copas passadas, clubes em que os jogadores das seleções atuam, desempenho desses jogadores em outras competições e assim por diante.

O formato desses multiplicadores varia de acordo com o site, e cada país têm geralmente um formato predominante. A seguir, seguem os formatos existentes atualmente:

2.3.1 *Odds* fracionais

Esse formato consiste no uso de frações para indicar o lucro caso sua aposta seja vencedora. Por exemplo, se a Suíça tiver uma *odds* de ganhar determinada partida de 8/1, indica que para cada real apostado, o apostador tem 8 reais de lucro, ou seja, recebe um pagamento total de 9 reais. Matematicamente temos

$$R = A \times \frac{N}{D} + A, \tag{2.22}$$

em que R é o retorno ou pagamento ou total recebido, A é o valor apostado, $\frac{N}{D} = O_f$ são as *odds* fracionais, N é o numerador da fração e D o denominador. Em palavras, caso o apostador ganhe, ele recebe o valor apostado A vezes o multiplicador $(\frac{N}{D})$ mais o valor inicial apostado. No exemplo acima, se o apostador tivesse apostado 32 reais e a Suíça vencesse sua partida, ele teria um retorno de $32 * \frac{8}{1} + 32 = 288$ reais.

2.3.2 *Odds* decimais

Neste formato, o cálculo do retorno R é facilitado, pois o multiplicador é aplicado diretamente sobre o valor apostado, e o resultado dessa multiplicação é o total recebido pelo apostador. Por exemplo, se o Brasil tem uma *odds* decimal de 1.46 contra a Sérvia, indica que se ele apostasse 257 reais no Brasil, ele recebe um total de $257 \times 1.46 = 375.22$, ou seja, obteve um lucro de $375.22 - 257 = 118.22$ reais. Matematicamente:

$$R = O_d \times A \tag{2.23}$$

em que R e A têm o mesmo significado que em (2.22) e O_d é a *odds* decimal.

2.3.3 *Odds Moneyline*

Nesta modalidade, aplicada geralmente a confrontos (um time contra o outro no caso do futebol), o favorito recebe um sinal (-) antes do valor da *odds* e a "zebra" recebe um sinal positivo (+). As *odds* de cada um desses times têm interpretações diferentes: no caso do favorito, indica o valor que o apostador deve colocar para lucrar 100 reais, enquanto que para a "zebra" indica o lucro a cada 100 reais apostados. Suponhamos o confronto entre Argentina e México, em que temos as seguintes *odds Moneyline* fictícias:

- Argentina: -829,
- México: $+640$.

Isso indica que para o apostador lucrar 100 reais apostando na Argentina, ele precisa desembolsar 829 reais. Caso ele opte por apostar no México, ele terá um lucro de 640 reais para cada 100 reais apostados. Matematicamente:

$$R = \begin{cases} \frac{100 \times A}{O_{m-}} + A, & \text{se (-)} \\ \frac{A \times O_{m+}}{100} + A, & \text{se (+)} \end{cases} \tag{2.24}$$

em que R e A têm a mesma definição indicada na equação (2.22), O_{m-} é o valor da *odds moneyline* para o favorito e O_{m+} é o mesmo valor para a "zebra".

2.4 Convertendo *odds* em probabilidades

Para que as *odds* sejam incorporadas ao modelo probabilístico proposto neste trabalho, é necessário transformá-las em probabilidades.

Quando os sites de aposta esportiva decidem pelo valor das *odds*, na verdade estão estabelecendo probabilidades de ocorrência do evento a ser apostado. Intuitivamente, temos que os maiores pagamentos por valor apostado pertencem a eventos com menor chance de acontecerem (ou seja, com menor probabilidade). Em outras palavras, o apostador corre um risco de perda maior no evento em que apostou, porém caso acerte sua recompensa também é maior.

Para cada tipo de *odds* (fracional, decimal e *moneyline*), temos um método diferente de conversão. A descrição de cada método é feita a seguir, utilizando a mesma notação adotada na seção 2.3, em que P é a probabilidade obtida a partir de determinada *odds*.

2.4.1 Conversão de *odds* fracionais em probabilidades.

A conversão das *odds* fracionais $O_f = \frac{N}{D}$ na probabilidade P é dada a seguir:

$$P = \frac{D}{N + D}, \tag{2.25}$$

em que a probabilidade P é dada pelo numerador de O_f dividido pela soma do numerador e denominador.

2.4.2 Conversão de *odds* decimais em probabilidades.

A conversão de uma *odds* decimal O_d em probabilidade é dada pelo inverso de O_d:

$$P = \frac{1}{O_d}, \tag{2.26}$$

2.4.3 Conversão de *odds moneyline* em probabilidades.

No caso das *odds moneyline*, o método de conversão depende se estamos lidando com o time favorito (-) ou seu adversário (+):

$$P = \begin{cases} \frac{O_{m-}}{O_{m-}+100}, & \text{se (-)} \\ \frac{100}{O_{m+}+100}, & \text{se (+)} \end{cases} \tag{2.27}$$

2.5 Avaliação do Modelo

Dado o resultado de uma partida, é necessário estabelecer métricas sobre a qualidade das previsões feitas pelo modelo. Para isso, usaremos as chamadas regras de escore (*scoring rules*), que avaliam a distância de predições probabilísticas de determinado resultado a aquele que de fato ocorreu. No caso deste trabalho, as predições podem ser representadas pelo vetor $P = (P_A, P_E, P_B)$, em que P_A é a probabilidade de vitória do time A, P_B a probabilidade de vitória do time B e P_E a de empate. Assim que determinado resultado é observado, ele será representado por $(1, 0, 0)$, $(0, 0, 1)$ ou $(0, 1, 0)$, se o time A ganhar, se o time B ganhar ou se o jogo terminar empatado, respectivamente. O vetor desses possíveis resultados será dado por $R = \{A, E, B\}$

2.5.1 Escore logarítmico

Dado um resultado $r \in R$ e um vetor de predições P, o escore logarítmico é dado por

$$S(P, r) = -\sum_{i \in R} \mathbb{1}(r = i) \ln(P_i), \tag{2.28}$$

que é o negativo da log-verossimilhança do evento ocorrido. Notemos que a predição "trivial" é dada por $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$, que produz um escore logarítmico de 1.098. Portanto, buscaremos predições que sejam pelo menos menores que esse valor, mais especificamente no intervalo $[0, 1.098)$, em que 0 corresponde à predição perfeita (por exemplo prever o vetor (1,0,0) e em seguida o time A ganhar).

2.5.2 Escore esférico

O escore esférico é dado por

$$S(P, r) = -\frac{1}{\sqrt{\sum_{i \in R} P_i^2}} \sum_{i \in R} \mathbb{1}(r = i) P_i, \tag{2.29}$$

que é o negativo da probabilidade do evento ocorrido, normalizado pela raiz da soma dos quadrados das probabilidades. Serão buscadas previsões no intervalo $[-1, -0.557)$, em que -1 corresponde à predição perfeita e -0.557 à trivial

2.5.3 Escore de Brier

O escore de Brier é calculado da seguinte forma:

$$S(P, r) = \sum_{i \in R} \mathbb{1}(r = i)(1 - P_i)^2 + \sum_{i \in R} \mathbb{1}(r \neq i) P_i^2 \tag{2.30}$$

Notemos que a predição "trivial" produz um escore de Brier igual a $\frac{2}{3}$, independentemente do resultado da partida. O intervalo de predição aceitável nesse caso é $[0, 0.667)$.

Capítulo 3

Resultados

Os resultados estão divididos em três seções: a seção 3.1 contém uma Análise descritiva. A seção 3.2 trata das três simulações feitas antes de cada uma das três rodadas da fase de grupos, sendo denominadas Simulação 1, 2 e 3. Por fim, a seção 3.3 trata das Simulações 4, 5 6 e 7, feitas antes de cada uma das rodadas das fases eliminatórias (oitavas, quartas, semis e final com disputa de 3º lugar).

Cada Simulação na verdade é um conjunto de 50000 simulações de todo o torneio a partir da rodada a que se referem. A Simulação 1 portanto tenta simular todo o torneio 50000 vezes antes que qualquer partida tenha sido observada. Sem jogos observados, o modelo preditivo "a priori" das equações (2.7) e (2.8) é usado para tentar prever, jogo a jogo, os resultados de todo o torneio. Após todos os jogos da primeira rodada da Fase de Grupos, suas informações são incorporadas ao modelo para tentar prever o restante da competição a partir da segunda rodada, que é a Simulação 2. A partir desta até a sétima Simulação é utilizado o modelo "a posteriori" representado por (2.12).

3.1 Análise Descritiva

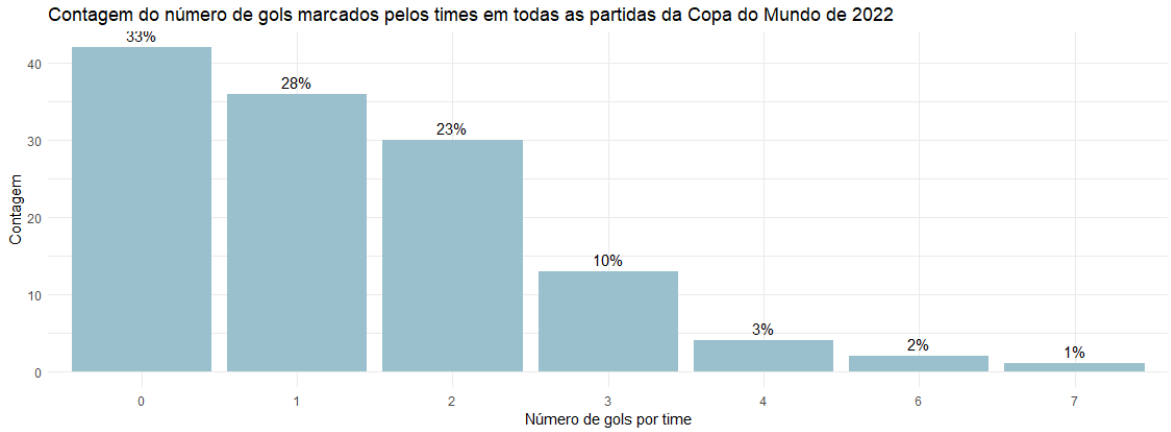

Figura 3.1: Frequência observada de número de gols marcados, por time, em todas as partidas disputadas na Copa do Mundo.

É possível perceber pela Figura 3.1 que a contagem do número de gols decresce à medida que o número de gols aumenta. Em outras palavras, pode-se dizer que partidas com alto número de gols marcados por um mesmo time são mais improváveis de acontecer que partidas com um menor número de gols. Particularmente, partidas com zero um ou dois gols marcados pelo mesmo time em uma partida representam 84% das ocorrências.

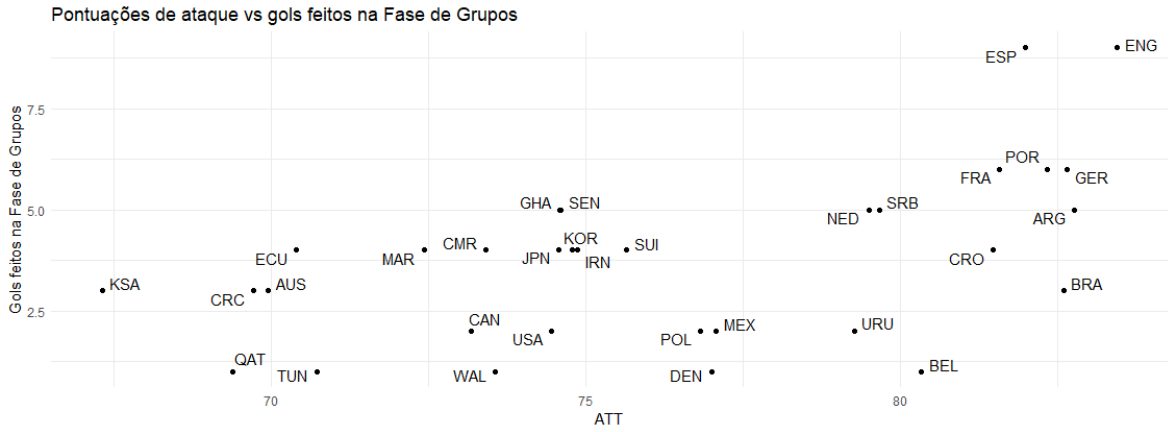

Figura 3.2: Cada ponto representa conjuntamente a pontuação de ataque do time (eixo das abscissas) e o número total de gols que esse time marcou na Fase de Grupos (eixo das ordenadas).

O padrão dos pontos representados na Figura 3.2 parece indicar alguma tendência de crescimento, ou seja, times com maior pontuação de ataque, particularmente com um valor dessa pontuação maior que 80, tendem a marcar mais gols em suas partidas da Fase de Grupos. Isso indica alguma correspondência entre as pontuações de *ATT* consideradas e

o desempenho do ataque do times durante a competição. É interessante notar porém que para escores menores que 80 essa tendência torna-se menos evidente, ainda que presente. Apenas partidas desta etapa foram consideradas (ao invés de considerar também as partidas das fases eliminatórias) pois os times jogam uma mesma quantidade de partidas em um cenário que todos os times ainda estão na competição. Acreditamos que desta forma seria possível uma comparação mais "justa" entre as equipes.

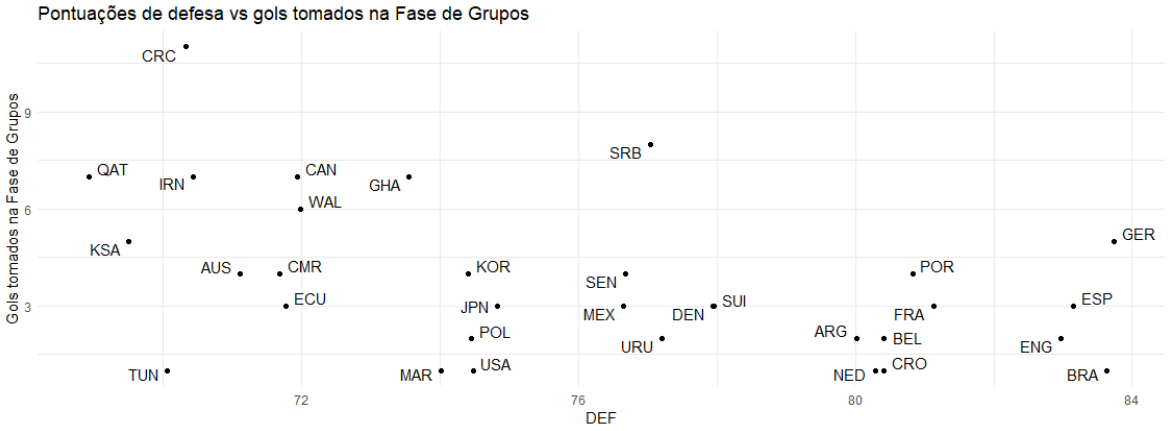

Figura 3.3: Cada ponto representa conjuntamente a pontuação de defesa do time (eixo das abscissas) e o número total de gols que o time sofreu na Fase de Grupos (eixo das ordenadas).

Constatação similar ao da Figura 3.2 pode ser observada na Figura 3.3. Nesta Figura, times com pontuação de defesa menor que 72 tendem a sofrer um maior número de gols, o que indica alguma correspondência entre as pontuações de DEF e o desempenho da defesa dos times observados durante a competição. Assim como no caso das pontuações de ATT, porém, essa tendência tende a ser menos evidente quando a pontuação de DEF é maior que 72, ainda que seja possível notá-la.

3.2 Simulações 1, 2 e 3: simulando o torneio antes de cada uma das rodadas da Fase de Grupos

As simulações foram feitas para todos os grupos. Para ilustrar o trabalho, porém, foi escolhido comentar apenas sobre o grupo do Brasil, ou seja, o Grupo G. Essa escolha foi feita pois o interesse nacional no desempenho do Brasil tende a ser maior, mas análises similares podem ser conduzidas para quaisquer outros grupos. Os gráficos referentes a esses outros grupos estão ilustrados no Apêndice A.6.

3.2.1 Grupo G

Probabilidades de Classificação dos times do Grupo G após término da Fase de Grupos

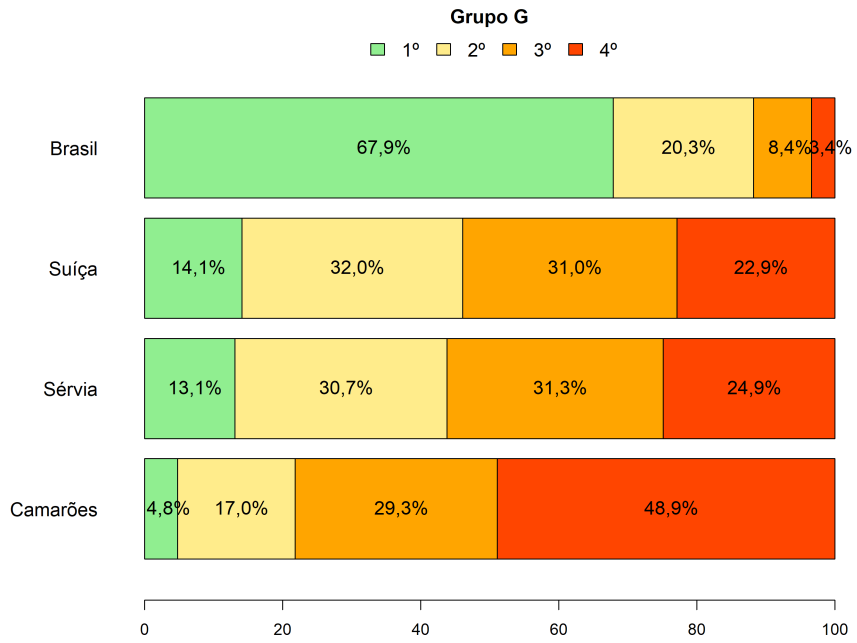

Figura 3.4: Simulação 1 (antes do início da Copa): probabilidades de cada time do grupo G terminar em determinada posição dentro do grupo após a simulação da disputa dos jogos das três rodadas da Fase de Grupos.

A Figura 3.4 retrata a simulação das três rodadas da Fase de Grupos para o Grupo G antes do início da competição. Nota-se o favoritismo do Brasil em terminar na primeira colocação (62%) do grupo. Tal favoritismo deve-se ao fato do Brasil ter as melhores pontuações de ATT e DEF entre os times do grupo, além das maiores *odds* de vitória. Lembrando que a Simulação 1 utiliza somente as informações provenientes da pontuação dos jogadores no jogo FIFA 23 e a expectativa das casas de aposta, dado que ainda não ocorreram jogos. O mesmo raciocínio vale para os outros times do grupo, com a Sérvia e Suíça disputando pela segunda colocação e Camarões como provável último colocado.

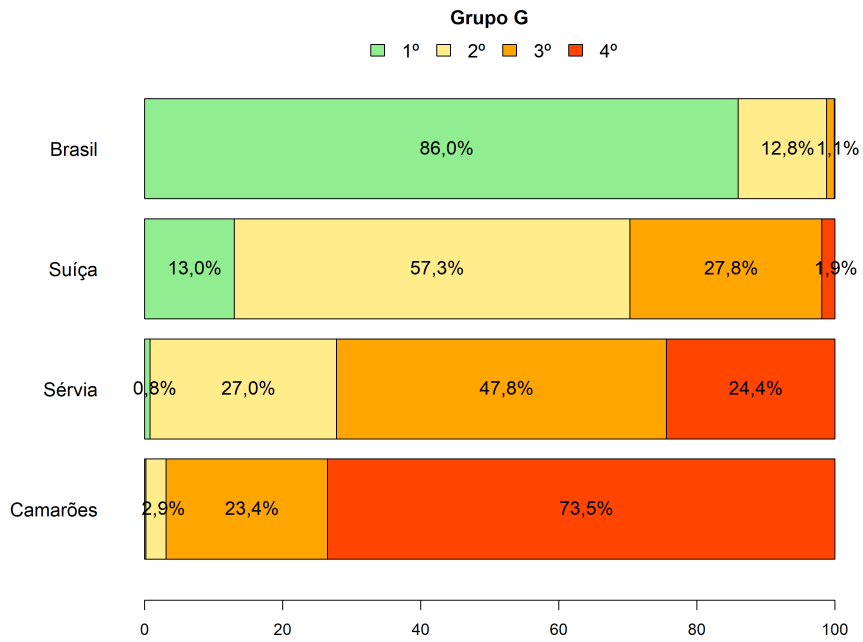

Figura 3.5: Simulação 2 (após o primeiro jogo de cada time): probabilidades de cada time do grupo G terminar em determinada posição dentro do grupo após a simulação da disputa dos jogos da segunda e terceira rodada da Fase de Grupos.

Após a realização dos primeiros jogos de cada time, foi feita a Simulação 2, com a Figura 3.5 representando as prováveis colocações após as simulações do segundo e terceiro jogos de cada time na Fase de Grupos. O Brasil aumentou seu favoritismo como primeiro colocado do grupo, pois além de possuir maiores pontuações de ATT e DEF e *odds* favoráveis, ganhou de 2 x 0 da Sérvia em seu primeiro jogo. A Sérvia, que tinha probabilidade similar a da Suíça de terminar em segundo lugar na Simulação 1, passa a tornar-se um provável terceiro colocado, dado que além de perder para o Brasil, a Suíça ganhou de Camarões por 1 x 0. Camarões portanto aumenta sua probabilidade de terminar como último colocado do time, dado que além dos piores pontuações de ATT e DEF e *odds* desfavoráveis, perdeu seu primeiro jogo sem ter feito nenhum gol (lembremos que neste modelo o número de gols do time o favorece em seus próximos jogos).

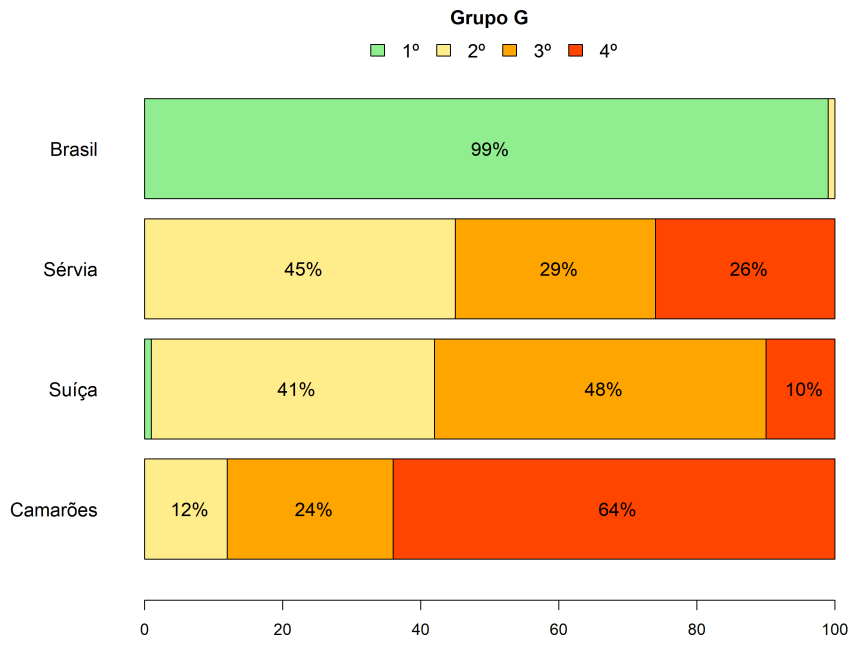

Figura 3.6: Simulação 3 (após o segundo jogo de cada time): probabilidades de cada time do grupo G terminar em determinada posição dentro do grupo após a simulação da disputa dos jogos da terceira rodada da Fase de Grupos.

Após o segundo jogo de cada time, foi feita a Simulação 3, que simula a competição a partir último jogo de cada time na Fase de Grupos. Percebe-se pela Figura 3.12 que o Brasil está praticamente garantido como primeiro colocado. Isso deve-se principalmente ao fato de que o Brasil foi o único time do grupo que ganhou seus dois primeiros jogos, vencendo a Suíça no segundo jogo por 1 x 0. Mais especificamente, mesmo se o Brasil perdesse seu último confronto contra Camarões (o que acabou se concretizando), apenas uma vitória da Suíça com um número de gols maior que 2 em relação à partida do Brasil possibilitaria a Suíça classificar-se em primeiro lugar. Em relação ao modelo proposto, nota-se ainda outros dois fatores que dificultam ainda mais a liderança da Suíça: Camarões, adversário do Brasil, é o time com as piores pontuações do grupo, o que favorece a probabilidade de vitória brasileira, e a pontuação de ATT da Suíça é menor que a pontuação de DEF da Sérvia, o que prejudica sua chance de marcar um grande número de gols a partir do modelo.

A Suíça passa então a disputar pelo 2º lugar com Sérvia e Camarões, ainda que o time africano tenha menor probabilidade de assumir essa posição. A vantagem que a Suíça tinha em assumir a segunda colocação na Simulação 2 diminuiu devido a sua derrota para o Brasil, porém não foi tão acentuada pois a Sérvia empatou com Camarões.

Após a terceira rodada o Brasil ficou em 1º lugar, com 99% de chance de terminar nesta

colocação segundo o modelo mais atualizado, ou seja, da Simulação 3; Suíça em segundo lugar, com 41%; Camarões (3º, 24%) e Sérvia (4º, 26%). Nota-se que a colocação mais surpreendente segundo essas probabilidades pertence a Camarões, que garantiu a terceira posição com uma vitória diante do Brasil por 1 x 0.

Performance do modelo nos jogos do Grupo G

As tabelas a seguir contêm o cálculo do escores de cada partida do grupo G, calculados a partir das Simulações 1, 2 e 3. Os maiores escores pertencem às partidas em que acontecem "'zebras", ou seja, partidas em que o time considerado não favorito pelo modelo vence. De forma recíproca, os menores escores pertencem às partidas em que o time considerado favorito vence. Sendo assim, escores menores indicam que o modelo tem melhor desempenho pelo critério adotado.

Tabela 3.1: Escores logarítmico, Esférico e de Brier para as partidas na Simulação 1.

			Probabilidades			Resultado Real			Escores		
Rodada	Time A	Time B	Vitória A	Empate	Vitória B	A ganhou	Empate	B ganhou	Logarítmico	Esférico	de Brier
1	BRA	SRB	0,63	0,19	0,18	1	0	0	0,46	-0,9	0,21
1	SUI	CMR	0,41	0,2	0,39	1	0	0	0,89	-0,68	0,54
2	BRA	SUI	0,68	0,15	0,17	1	0	0	0,39	-0,93	0,15
2	CMR	SRB	0,12	0,47	0,41	0	1	0	0,89	-0,81	0,58
3	SRB	SUI	0,33	0,36	0,31	0	0	1	1,02	-0,65	0,61
3	CMR	BRA	0,07	0,13	0,8	1	0	0	2,66	-0,1	1,52

Da tabela 3.1 temos a simulação das três partidas de cada time do Grupo G antes do início da Copa do Mundo, ou seja, o modelo dado pela preditiva a priori (equações 2.7 e 2.8). Nesta simulação, exceto por dois casos, os eventos com maior probabilidade predita foram os que de fato ocorreram, inclusive no empate entre Camarões e Sérvia. A primeira exceção ocorreu na partida entre Suíça e Camarões, porém neste caso a diferença entre as probabilidades era menor, ou seja, sem um evento claramente favorito.

A exceção mais marcante ocorreu na partida entre Brasil e Camarões, o que é de se esperar dadas as diferenças entre suas pontuações de ATT e DEF e as *odds* da partida (Tabela A.1).

Tabela 3.2: Escores logarítmico, Esférico e de Brier para as partidas na Simulação 2.

			Probabilidades			Resultado Real			Escores		
Rodada	Time A	Time B	Vitória A	Empate	Vitória B	A ganhou	Empate	B ganhou	Logarítmico	Esférico	de Brier
2	BRA	SUI	0,68	0,18	0,14	1	0	0	0,39	-0,93	0,16
2	CMR	SRB	0,14	0,46	0,4	0	1	0	0,78	-0,89	0,48
3	SRB	SUI	0,27	0,37	0,36	0	0	1	1,02	-0,67	0,62
3	CMR	BRA	0,04	0,14	0,82	1	0	0	3,21	-0,06	1,61

Decorridos os primeiros jogos de cada time e incorporados os placares e pontuações

de ATT e DEF observados ao modelo, foi feita a Simulação 2, que simula a competição a partir do segundo jogo de cada time segundo a equação (2.12).

A partir da tabela 3.2 é possível comparar as simulações em relação à segunda e terceira rodadas. Da simulação 1 para a 2 houve um aumento de 5% na probabilidade de vitória do Brasil contra a Suíça, o que diminuiu o valor dos escores para esta partida. No outro jogo de segunda rodada entre Camarões e Sérvia o modelo manteve uma maior probabilidade de empate, que de fato veio a acontecer. Há indícios portanto que o modelo atribui de forma satisfatória as probabilidades das partidas da segunda rodada para este grupo.

Ao observar as previsões da terceira rodada, porém, nota-se aumento do valor do escore no jogo entre Brasil e Camarões, isto é, aumentou a probabilidade de vitória do Brasil, que acabou perdendo. Esse aumento ocorreu pois o Brasil confirmou seu favoritismo na primeira rodada, ganhando da Sérvia fazendo 2 gols, enquanto que Camarões perdeu para a Suíça sem fazer nenhum gol, reforçando sua posição como pior time do grupo. Os sites de aposta esportiva também ajustaram suas expectativas ainda mais a favor do Brasil, como é possível observar nas Tabelas A.1 e A.2 do Apêndice.

Em relação ao jogo de terceira rodada entre Sérvia e Suíça não houve tendência clara de aumento ou diminuição dos escores, dado que os três escores tiveram comportamentos diferentes.

Tabela 3.3: Escores logarítmico, Esférico e de Brier para as partidas na Simulação 3.

			Probabilidades			Resultado Real			Escores		
Rodada	Time A	Time B	Vitória A	Empate	Vitória B	A ganhou	Empate	B ganhou	Logarítmico	Esférico	de Brier
3	CMR	BRA	0,19	0,20	0,61	1	0	0	1,64	-0,32	1,06
3	SRB	SUI	0,52	0,28	0,20	0	0	1	1,61	-0,31	0,99

Da tabela 3.3 percebe-se que houve diminuição dos valores dos escores calculados para o jogo entre Brasil e Camarões. Isso pode ser explicado pelo jogo de Camarões contra a Sérvia na rodada anterior. Camarões, apesar de ter empatado, fez três gols, o que favorece a performance do time em sua(s) próxima(s) partida(s). Além disso, as *odds* da simulação 2 para uma vitória de Camarões era 19, enquanto que na simulação 3 passam a ser de 12,5, ou seja, os sites de aposta esportiva aumentaram as chances de Camarões contra o Brasil. Apesar de ainda ser um valor alto de *odds*, percebe-se que os sites de aposta acreditaram em um aumento das chances de vitória de Camarões após observarem os jogos da segunda rodada.

Performance do modelo em todos os jogos da Fase de Grupos

Para cada *boxplot* a seguir, a linha pontilhada na horizontal indica o valor de referência do escore, ou seja, o valor do escore no caso de atribuirmos uma probabilidade de $\frac{1}{3}$ para os possíveis resultados de uma partida na Fase de Grupos (vitória, empate e derrota). Sendo assim, valores abaixo do valor de referência indicam partidas em que o modelo performou melhor que o caso base, e valores acima indicam os casos em que essa performance foi pior.

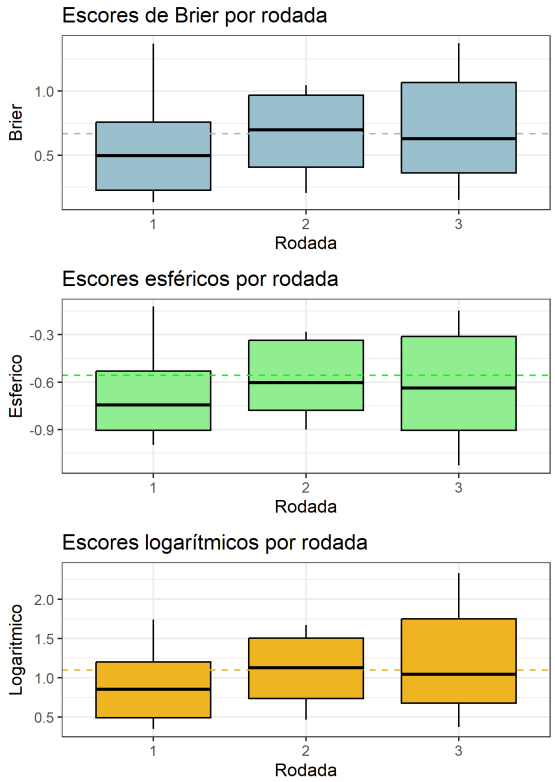

Figura 3.7: Distribuição dos escores após a simulação das três rodadas da fase de grupos (Simulação 1).

A partir da Figura 3.7 observa-se que o modelo tem performance decrescente no decorrer da Fase de Grupos, ou seja, possui capacidade de prever melhor a primeira rodada que a segunda, e a segunda melhor que a terceira. Esta tendência é percebida em todas as simulações da Fase de Grupos (1,2 e 3) e ilustrada nos *boxplots*. Este comportamento é esperado, particularmente na terceira rodada, em que algumas situações dificultam a atribuição de probabilidades. Por exemplo, há casos de times considerados "fortes" que podem já estar classificados antes de seu jogo de terceira rodada, e assim poupam seus esforços para a fase eliminatória, aumentando suas chances de derrota ou empate na terceira rodada. Enquanto isso, times que necessitam de uma vitória obrigatoriamente para

continuar na competição necessitam maximizar seu esforço para continuarem na competição.

Nota-se o modelo performou melhor que o caso base em aproximadamente 10 jogos dos 16 disputados na primeira rodada. As principais "zebras" desta rodada foram a vitória da Arábia Saudita sobre a Argentina, e do Japão sobre a Alemanha.

Tabela 3.4: Escore médio, escore base e diferença entre estes valores para os jogos da primeira rodada da Fase de Grupos na Simulação 1.

	Escore médio (1)	Escore base (2)	Diferença % (2-1)
Logarítmico	0,94	1,10	14%
Esférico	-0.68	-0,56	18%
Brier	0,54	0,67	19%

Percebe-se na Tabela 3.4 que o escore médio é menor que o escore base nos três casos. O modelo, que na Simulação 1 leva em conta apenas as informações das *odds* e das pontuações de ATT e DEF conseguiu escores abaixo do valor de referência (abaixo neste caso é algo positivo).

Tabela 3.5: Escore médio, escore base e diferença entre estes valores para os jogos da segunda rodada da Fase de Grupos na Simulação 1.

	Escore médio (1)	Escore base (2)	Diferença % (2-1)
Logarítmico	1,10	1,10	0%
Esférico	-0.59	-0,56	5%
Brier	0,66	0,67	1%

Na segunda rodada da Fase de Grupos, o modelo sem jogos observados obteve um escore médio idêntico ao valor de referência no caso do escore logarítmico, e um pouco acima no caso dos escores esférico e de Brier, indicando um declínio no poder preditivo em relação à primeira rodada.

Tabela 3.6: Escore médio, escore base e diferença entre estes valores para os jogos da terceira rodada da Fase de Grupos na Simulação 1.

	Escore médio (1)	Escore base (2)	Diferença % (2-1)
Logarítmico	1,16	1,10	-5%
Esférico	-0.64	-0,56	-12%
Brier	0,68	0,67	-1%

Nos 16 jogos da última rodada da Fase de Grupos, o modelo sem jogos observados obteve uma queda de performance, o que já era esperado pelo fenômeno descrito no parágrafo abaixo da Figura 3.7.

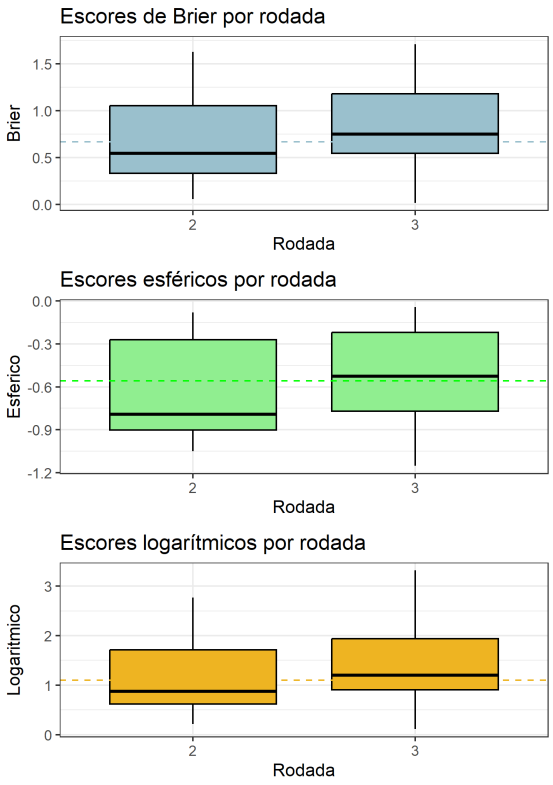

Figura 3.8: Distribuição dos escores após a simulação das duas últimas rodadas da fase de grupos (Simulação 2).

Na Figura 3.8 é possível notar um comportamento similar ao da Figura 3.5. Há uma menor precisão na previsão dos resultados da terceira rodada da Fase de Grupos em comparação a rodada anterior.

Tabela 3.7: Escore médio, escore base e diferença entre estes valores para os jogos da segunda rodada da Fase de Grupos na Simulação 2.

	Escore médio (1)	Escore base (2)	Diferença % (2-1)
Logarítmico	1,21	1,10	-9%
Esférico	-0.61	-0,56	-8%
Brier	0,71	0,67	-6%

Da Tabela 3.7 percebe-se que os escores do modelo na Simulação 2, ou seja, após acontecerem os jogos da primeira rodada da Fase de Grupos, é ligeiramente abaixo do escore base. Houve uma queda do poder preditivo nesta rodada em comparação a Simulação 1 (modelo sem jogos), ilustrado na Tabela 3.5. Este comportamento pode ser explicado por algumas sequências de jogos atípicos: Marrocos, um dos times "revelação" da competição, conseguindo terminar em 4º lugar, em seu jogo de primeira rodada empatou com a Croácia sem ter feito nenhum gol, o que não favorece suas chances na partida contra a Bélgica, um time considerado mais "forte" pelo modelo. Porém na segunda rodada o time

marroquino ganhou da Bélgica, obtendo um escore logarítmico de 1,75. Isso é aumento em relação a Simulação 1 para este jogo, que teve um escore de 1,64. Porém os principais aumentos ocorreram devido ao desempenho do Japão e da Inglaterra. No caso do Japão, na Simulação 1, o escore logarítmico para o jogo de segunda rodada contra a Costa Rica era de 1,67, um escore considerado alto pois ele era favorito mas perdeu. Porém na primeira rodada o time foi capaz de ganhar da Alemanha, considerada forte candidata a atingir estágios avançados da competição. Sendo assim, na Simulação 2, que considerou esta vitória antes do jogo contra a Costa Rica, aumentou ainda mais a probabilidade de vitória do Japão (42% na simulação 1 contra 73% na Simulação 2). Com a derrota para a Costa Rica, o escore logarítmico desta partida aumentou para 2,76 (contra 1,67 da Simulação 1). Situação similar ocorreu no jogo de segunda rodada entre Inglaterra e Estados Unidos. A Inglaterra ganhou seu primeiro jogo contra o Irã por 6 x 2. O alto número de gols da Inglaterra no primeiro jogo aumentou sua probabilidade de vitória na partida contra os Estados Unidos, que vinha de empate contra o País de Gales por 1 x 1. A probabilidade de vitória da Inglaterra contra os Estados Unidos era de 58% na Simulação 1 (antes de ganhar do Irã), mas passou a ser de 87% na Simulação 2 (após ganhar do Irã tendo feito 6 gols). Neste jogo, porém, a Inglaterra apenas empatou com os Estados Unidos, gerando um escore logarítmico de 2,60 para esta partida na Siulação 2 (versus 1,48 na Simulação 1).

Tabela 3.8: Escore médio, escore base e diferença entre estes valores para os jogos da terceira rodada da Fase de Grupos na Simulação 2.

	Escore médio (1)	Escore base (2)	Diferença % (2-1)
Logarítmico	1,46	1,10	-25%
Esférico	-0.53	-0,56	-5%
Brier	0,83	0,67	-19%

Os jogos de terceira rodada na Simulação 2, assim como na Simulação 1, apresentam desempenho abaixo do valor de referência, fenômeno explicado no parágrafo abaixo da Figura 3.7.

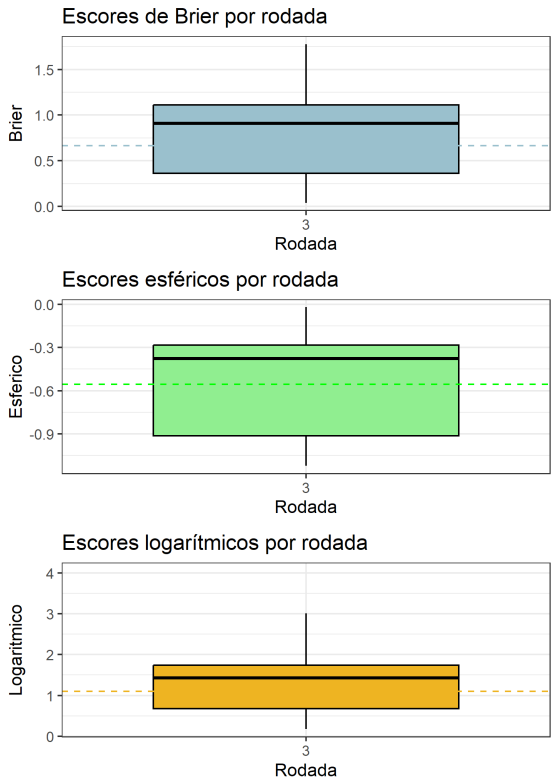

Figura 3.9: Distribuição dos escores após a simulação da última rodada da fase de grupos (Simulação 3).

Na Figura 3.9 é possível notar o mesmo comportamento da Figura 3.8, com desempenho inferior ao escore base em mais da metade dos jogos. Nota-se que a incorporação do número de gols ao modelo desfavoreceu sua capacidade preditiva para os jogos da terceira rodada, que já tende a ser de difícil previsão.

Tabela 3.9: Escore médio, escore base e diferença entre estes valores para os jogos da terceira rodada da Fase de Grupos na Simulação 3.

	Escore médio (1)	Escore base (2)	Diferença % (2-1)
Logarítmico	1,45	1,10	-24%
Esférico	-0.53	-0,56	-5%
Brier	0,82	0,67	-18%

Nota-se que na Simulação 3 os escores médios ficaram acima dos escores base, indicando uma performance pior que a referência adotada, particularmente visível pelo escore logarítmico.

3.2.2 Simulação de todo o torneio para os 16 melhores times após término da Copa do Mundo.

As tabelas a seguir indicam, para os 16 melhores classificados após o término da Copa do Mundo, suas probabilidades de atingir cada etapa do torneio nas simulações 1, 2 e 3. A simulação 1 (Tabela 3.10) refere-se a simulação de todo o torneio com os dados disponíveis antes da primeira rodada da Fase de Grupos, usando o Modelo Sem Jogos apresentado na subseção 2.1.1. A simulação 2 (Tabela 3.11) também simula de todo o torneio, porém a partir da segunda rodada da Fase de Grupos, ou seja, após os jogos da primeira rodada ocorrerem. O mesmo vale para a simulação 3, que simula o torneio antes dos jogos da terceira rodada. Note que as simulações 2 e e 3 utilizam-se do Modelo com Jogos descrito na subseção 2.1.2.

Probabilidades de cada time atingir determinada rodada do torneio

Tabela 3.10: Probabilidades de cada time atingir determinada rodada do torneio, de acordo com a simulação antes dos primeiros jogos da Fase de Grupos (simulação 1).

				Probabilidades preditas pelo modelo em %								
				Fase de grupos				fases eliminatórias				
Sigla	Posição após Copa	Time	Grupo	Pos1	Pos2	Pos3	Pos4	Oitavas	Quartas	Semi	Final	Campeão
ARG	1	Argentina	C	63,14	23,03	9,83	4,00	86,17	55,92	35,45	19,33	10,64
FRA	2	França	D	60,05	26,26	9,84	3,85	86,31	57,53	36,30	20,62	11,20
CRO	3	Croácia	F	28,68	31,51	23,75	16,05	60,19	25,65	11,64	4,99	1,93
MAR	4	Marrocos	F	11,13	21,21	30,50	37,15	32,34	8,89	2,62	0,77	0,21
BRA	5	Brasil	G	67,94	20,27	8,43	3,37	88,21	65,07	41,16	26,12	16,46
ENG	5	Inglaterra	B	63,13	22,00	10,14	4,73	85,13	58,17	33,29	18,19	9,46
NED	5	Holanda	A	63,48	21,85	9,98	4,70	85,33	58,36	33,64	17,67	9,47
POR	5	Portugal	H	53,21	26,10	13,37	7,32	79,32	46,67	24,69	13,23	6,30
AUS	6	Austrália	D	5,42	17,56	35,45	41,57	22,98	6,06	1,58	0,32	0,05
JPN	6	Japão	E	7,96	21,10	41,84	29,11	29,05	10,52	3,53	1,06	0,28
KOR	6	Coreia do Sul	H	10,11	21,35	31,75	36,79	31,46	8,77	2,32	0,62	0,13
USA	6	Estados Unidos	B	16,69	31,37	28,61	23,32	48,07	19,28	6,40	1,91	0,44
ESP	7	Espanha	E	46,11	34,52	14,49	4,88	80,63	52,92	31,12	17,89	9,64
POL	7	Polônia	C	16,86	31,70	30,65	20,79	48,56	18,51	7,59	2,53	0,75
SEN	7	Senegal	A	15,63	30,30	29,82	24,25	45,93	18,41	6,15	1,96	0,58
SUI	7	Suíça	G	14,15	31,99	30,97	22,88	46,14	20,56	7,50	2,73	0,86

Como não há partidas observadas na simulação 1, as probabilidades representam as expectativas das casas de aposta e as pontuações dos jogadores no jogo FIFA 23. Portanto é natural, nesta simulação, que os times considerados de maior tradição em Copas do Mundo tenham maiores probabilidades de atingirem estágios mais avançados na competição.

Tabela 3.11: Probabilidades de cada time atingir determinada rodada do torneio, de acordo com a simulação antes dos jogos da segunda rodada da Fase de Grupos (simulação 2).

				Probabilidades preditas pelo modelo em %								
				Fase de grupos				fases eliminatórias				
Sigla	Posição após Copa	Time	Grupo	Pos1	Pos2	Pos3	Pos4	Oitavas	Quartas	Semi	Final	Campeão
ARG	1	Argentina	C	12,20	48,29	22,96	16,55	60,49	34,42	19,42	10,73	5,31
FRA	2	França	D	96,55	3,04	0,29	0,12	99,60	62,56	34,26	19,54	9,41
CRO	3	Croácia	F	12,11	30,69	33,95	23,25	42,80	16,85	6,27	2,69	0,96
MAR	4	Marrocos	F	11,21	27,73	32,73	28,33	38,94	12,20	3,50	1,17	0,34
BRA	5	Brasil	G	86,00	12,76	1,11	0,13	98,76	80,24	54,65	39,05	26,10
ENG	5	Inglaterra	B	96,51	3,10	0,27	0,12	99,61	69,58	38,22	22,70	11,28
NED	5	Holanda	A	76,92	21,49	1,44	0,14	98,41	71,14	51,91	25,66	14,09
POR	5	Portugal	H	92,08	6,49	0,98	0,45	98,57	60,89	36,15	17,47	7,60
AUS	6	Austrália	D	0,42	46,03	27,20	26,34	46,46	20,73	5,38	0,94	0,21
JPN	6	Japão	E	9,50	78,98	10,89	0,63	88,48	36,39	12,89	3,91	1,11
KOR	6	Coreia do Sul	H	1,62	13,75	36,88	47,75	15,37	3,29	1,00	0,32	0,07
USA	6	Estados Unidos	B	1,81	28,89	35,29	34,01	30,70	10,62	5,51	1,50	0,45
ESP	7	Espanha	E	89,96	8,51	1,51	0,03	98,47	70,23	35,21	22,40	12,41
POL	7	Polônia	C	6,04	16,69	34,92	42,35	22,73	7,87	2,44	0,68	0,19
SEN	7	Senegal	A	0,66	18,73	58,45	22,16	19,39	6,60	2,48	1,00	0,33
SUI	7	Suíça	G	13,01	57,28	27,75	1,96	70,29	24,28	8,67	2,58	0,64

Após o primeiro jogo, nota-se uma intensificação das probabilidades de avanço do Brasil. Isso pode ser explicado pela vitória do Brasil em seu primeiro jogo contra a Sérvia, além da derrota da Argentina para a Arábia Saudita. O time de Marrocos também empatou seu jogo contra a Croácia, e portanto ainda não dava indícios de que seria uma das surpresas da competição. Inglaterra, Holanda e Portugal aumentaram suas probabilidades de avançarem na competição pois venceram suas partidas de primeira rodada. O Japão foi um outro time que se destacou em seu primeiro jogo, ganhando da Alemanha, aumentando sua probabilidade de classificação para as oitavas de final de 29% na Simulação 1 (Tabela 3.10 para 88% na Simulação 2 (Tabela 3.11).

Tabela 3.12: Probabilidades de cada time atingir determinada rodada do torneio, de acordo com a simulação antes dos jogos da terceira rodada da Fase de Grupos (simulação 3).

				Probabilidades preditas pelo modelo em %								
				Fase de grupos				fases eliminatórias				
Sigla	Posição após Copa	Time	Grupo	Pos1	Pos2	Pos3	Pos4	Oitavas	Quartas	Semi	Final	Campeão
ARG	1	Argentina	C	54,42	16,47	28,99	0,13	70,88	47,95	26,41	11,82	5,13
FRA	2	França	D	99,99	0,01	0,00	0,00	100,00	72,14	36,23	15,29	6,56
CRO	3	Croácia	F	65,50	20,89	13,61	0,00	86,39	30,46	12,50	4,80	1,54
MAR	4	Marrocos	F	25,92	70,38	3,70	0,00	96,30	28,32	8,66	3,06	0,82
BRA	5	Brasil	G	98,77	1,23	0,00	0,00	100,00	83,58	52,85	34,27	19,26
ENG	5	Inglaterra	B	90,37	9,61	0,03	0,00	99,97	61,98	36,42	16,62	7,67
NED	5	Holanda	A	64,17	34,43	1,39	0,00	98,61	73,64	51,42	27,86	15,22
POR	5	Portugal	H	93,99	6,01	0,00	0,00	100,00	66,56	28,93	14,61	6,21
AUS	6	Austrália	D	0,01	61,19	38,21	0,60	61,19	18,71	4,20	0,56	0,10
JPN	6	Japão	E	4,91	2,98	89,80	2,32	7,88	2,52	0,53	0,08	0,01
KOR	6	Coreia do Sul	H	0,00	2,52	59,38	38,09	2,52	0,34	0,08	0,01	0,01
USA	6	Estados Unidos	B	1,18	20,89	71,57	6,35	22,08	7,85	3,51	0,95	0,26
ESP	7	Espanha	E	94,27	5,18	0,54	0,00	99,46	71,70	35,48	20,65	10,16
POL	7	Polônia	C	32,93	36,01	31,06	0,00	68,94	19,38	4,10	0,57	0,09
SEN	7	Senegal	A	10,19	28,48	60,86	0,47	38,67	12,48	4,28	0,90	0,18
SUI	7	Suíça	G	1,23	41,16	47,52	10,10	42,38	14,31	4,11	1,29	0,35

Após os jogos da segunda rodada, é feita a Simulação 3, com resultados representados na Tabela 3.12. Nota-se uma polarização das probabilidades da Fase de Grupos pois dois dos três jogos de cada time nesta etapa já foram disputados. Marrocos, que vem de empate na primeira rodada seguido de uma vitória sobre a Bélgica, praticamente garante sua classificação para as oitavas de final. É interessante notar que mesmo neste cenário sua probabilidade de atingir as fases posteriores cai drasticamente, ou seja, o modelo não considera a classificação de um time "zebra" às fases eliminatórias como um fator que eleva drasticamente seu desempenho nesta fase. O Japão, que vinha com probabilidade alta de classificação na Simulação 2 graças a sua vitória sobre a Alemanha, perde para a Costa Rica na segunda rodada, e assim o modelo aumenta sua probabilidade de derrota contra a Espanha na última rodada. A Espanha ganhou da Costa Rica por 7 x 0 na primeira rodada, o que favorece ainda mais o time em seu confronto contra o Japão.

Probabilidades de cada time ganhar o torneio

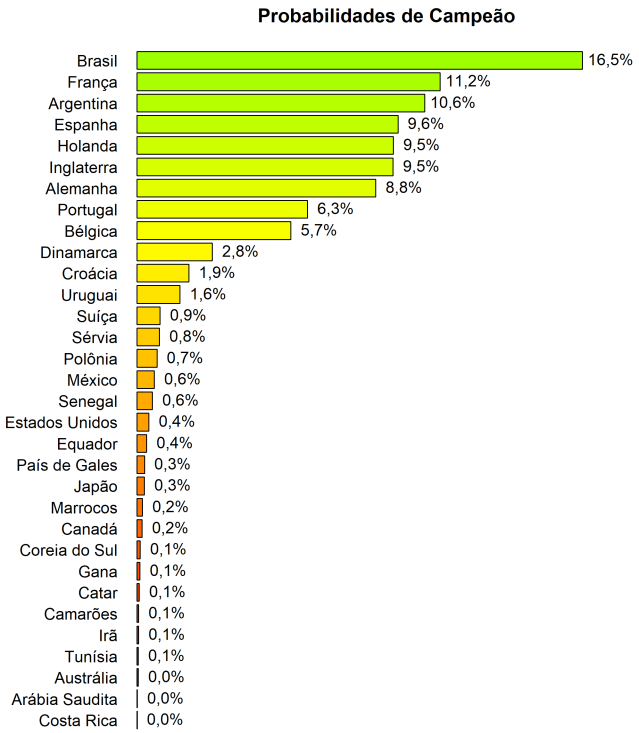

Figura 3.10: Probabilidades de campeão do torneio na Simulação 1

É possível perceber que na Simulação 1 o Brasil é considerado como favorito a ganhar a competição, com probabilidade de $16,5\%$, seguido da França ($11,2\%$) e Argentina ($10,6\%$). Lembremos que nesta etapa, como ainda não houve nenhuma partida, as pon-

tuações de ATT e DEF e as opiniões dos sites de aposta esportiva são as únicas fontes de informação do modelo, dado que não ocorreram jogos.

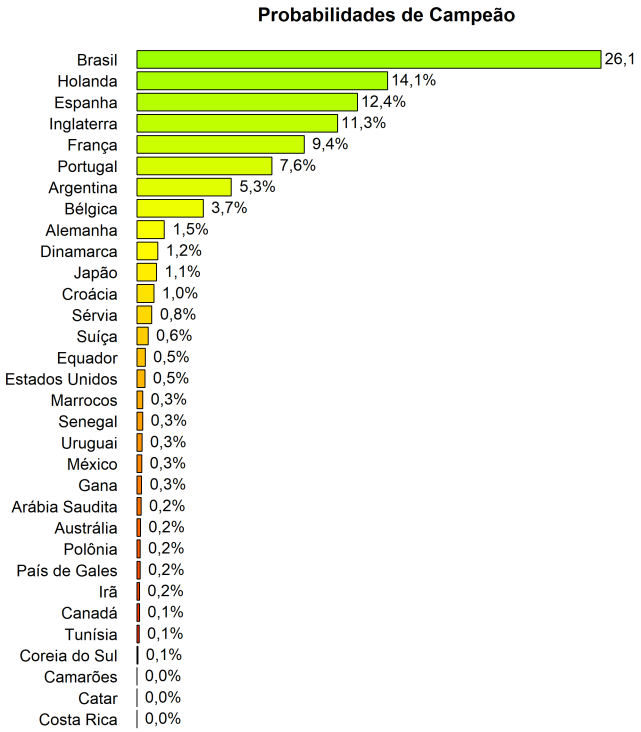

Figura 3.11: Probabilidades de campeão do torneio na Simulação 2

Após a primeira partida, aumenta-se a probabilidade do Brasil ser campeão dado a vitória do time na primeira rodada e diminui-se a probabilidade da Argentina, que perdeu sua primeira partida contra a Arábia Saudita. A goleada da Espanha sobre a Costa Rica por 7 x 0 também aumenta a probabilidade de campeã simulada nesta rodada. Esse resultado pode ter influenciado a diminuição da probabilidade da França mesmo tendo ganhado da Dinamarca, pois foi uma vitória com um placar menor (2 x 1).

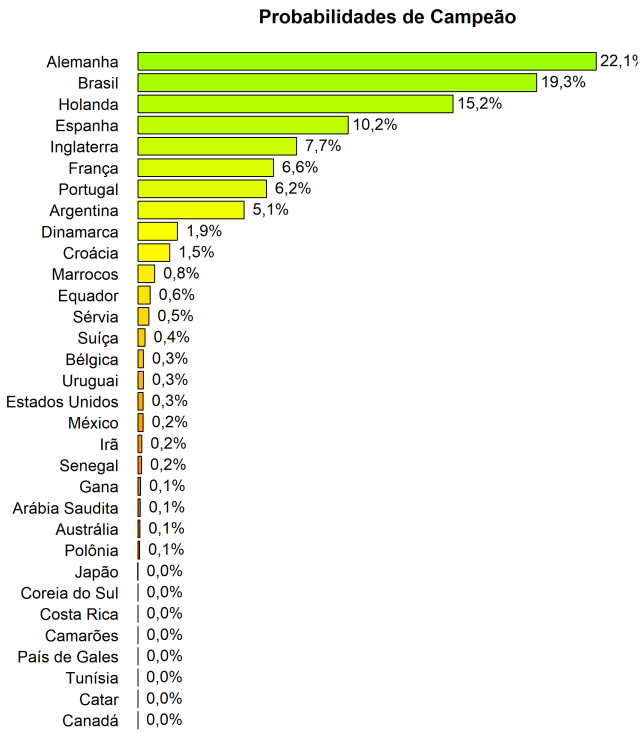

Figura 3.12: Probabilidades de campeão do torneio na Simulação 3

É interessante notar na Simulação 3 (após os jogos da segunda rodada) que a Alemanha é tida como favorita, mas a vitória do Japão contra a Espanha na terceira rodada fez com que a Alemanha terminasse na terceira colocação do grupo, não se classificando para as oitavas de final. O Brasil mantém seu favoritismo em ganhar a competição, tendo vencido seus dois primeiros jogos.

3.3 Simulações 4, 5, 6 e 7: simulando o torneio antes de cada uma das rodadas das fases eliminatórias (oitavas, quartas, semis e final).

A partir das oitavas de final, foram feitas as Simulações 4, 5, 6 e 7. A Simulação 4 simula o torneio a partir das oitavas de final até o jogo da final. A Simulação 5, a partir das quartas até a final (já com a informação dos resultados das oitavas) e assim por diante até a Simulação 7, referente à final e a disputa de 3º lugar.

A Tabela 3.13 abaixo contém os resultados da Simulação 4 para os jogos das oitavas, a Simulação 5 para os jogos das quartas, a Simulação 6 para os jogos das semis e a Simulação 7 para os jogos da final. Note que cada uma das simulações foi feita originalmente para

todas as etapas até a final, porém a Tabela 3.13 contém apenas os resultados em que a simulação tem os jogos definidos. Isso foi feito pois nas fases eliminatórias, ao contrário da Fase de Grupos, não é possível saber os confrontos das rodadas posteriores, e assim não é possível apostar nessas partidas. Sendo assim, o interesse maior é saber as probabilidades de cada simulação nos confrontos que irão de fato ocorrer.

Tabela 3.13: Probabilidades de vitória de cada time para cada uma das etapas das fases eliminatórias.

			Probabilidades		Resultado Real		Escores		
Rodada	Time A	Time B	Vitória A	Vitória B	A ganhou	B ganhou	Logarítmico	Esférico	de Brier
Oitavas	NED	USA	0,67	0,33	1	0	0,40	-0,90	0,21
Oitavas	ARG	AUS	0,93	0,07	1	0	0,07	-1,00	0,01
Oitavas	JPN	CRO	0,36	0,64	0	1	0,45	-0,87	0,26
Oitavas	BRA	KOR	0,91	0,09	1	0	0,09	-1,00	0,02
Oitavas	ENG	SEN	0,82	0,18	1	0	0,20	-0,98	0,06
Oitavas	FRA	POL	0,90	0,10	1	0	0,11	-0,99	0,02
Oitavas	MAR	ESP	0,20	0,80	1	0	1,61	-0,24	1,28
Oitavas	POR	SUI	0,69	0,31	1	0	0,37	-0,91	0,19
Quartas	NED	ARG	0,39	0,61	0	1	0,49	-0,84	0,30
Quartas	CRO	BRA	0,13	0,87	1	0	2,04	-0,15	1,51
Quartas	ENG	FRA	0,45	0,55	0	1	0,61	-0,77	0,41
Quartas	MAR	POR	0,23	0,77	1	0	1,45	-0,29	1,17
Semis	ARG	CRO	0,71	0,29	1	0	0,35	-0,92	0,17
Semis	FRA	MAR	0,82	0,18	1	0	0,19	-0,98	0,06
Disputa 3º lugar	CRO	MAR	0,57	0,43	1	0	0,56	-0,80	0,37
Final	ARG	FRA	0,51	0,49	1	0	0,68	-0,72	0,49

Nota-se na Tabela 3.13 que todos os escores, exceto por dois deles, performaram melhor que o valor de referência. Como não há a possibilidade de empate nesta etapa da competição, performar melhor que o valor de referência significa que o time considerado favorito pelo modelo de fato venceu sua partida. Sendo assim, apenas três "zebras" aconteceram: a vitória de Marrocos sobre a Espanha nas oitavas e sobre Portugal nas quartas, e a derrota do Brasil para a Croácia nas quartas.

Da Tabela 3.14 até a Tabela 3.17, cada Simulação (4,5,6 e 7) simulou todo o torneio, a partir das oitavas, das quartas, das semis e da final, respectivamente. Sendo assim, foi possível, para cada Simulação, obter probabilidades de cada time atingir determinada etapa das fases eliminatórias.

Tabela 3.14: Probabilidades de cada time atingir determinada etapa da competição nas fases eliminatórias na Simulação 4 (antes dos jogos das oitavas).

	Probabilidades			
Time	Quartas	Semis	Final	Campeão
Brasil	0,91	0,72	0,43	0,27
Argentina	0,93	0,72	0,40	0,25
França	0,90	0,56	0,37	0,18
Inglaterra	0,82	0,39	0,22	0,10
Espanha	0,80	0,49	0,22	0,09
Portugal	0,69	0,35	0,14	0,05
Holanda	0,68	0,21	0,07	0,03
Croácia	0,63	0,18	0,06	0,02
Suíça	0,31	0,10	0,02	0,00
Japão	0,37	0,07	0,02	0,00
Estados Unidos	0,32	0,06	0,01	0,00
Marrocos	0,20	0,06	0,01	0,00
Senegal	0,18	0,03	0,01	0,00
Coréia do Sul	0,09	0,03	0,00	0,00
Polônia	0,10	0,02	0,00	0,00
Austrália	0,07	0,02	0,00	0,00

Da Tabela 3.14 tem-se o Brasil e a Argentina como principais favoritos ao título, seguidos de França, Inglaterra e Espanha. Entre os times com as menores chances de avançarem às quartas, nota-se a Austrália (0,07), Coreia do Sul (0,09) e Polônia (0,10), que de fato viriam a perder seus jogos nas oitavas. É interessante notar que Marrocos, que ganhou nas oitavas da Espanha (um time "forte") tem uma probabilidade um pouco maior que os três times menos tradicionais supracitados, ainda que não seja tão expressiva (0,20).

Tabela 3.15: Probabilidades de cada time atingir determinada etapa da competição nas fases eliminatórias na Simulação 5 (antes dos jogos das quartas).

	Probabilidades		
Time	Semis	Final	Campeão
Brasil	0,87	0,66	0,48
Portugal	0,77	0,45	0,19
França	0,55	0,27	0,10
Argentina	0,61	0,20	0,10
Inglaterra	0,45	0,21	0,07
Holanda	0,39	0,09	0,04
Marrocos	0,23	0,07	0,01
Croácia	0,13	0,04	0,01

Após a vitória do Brasil contra a Coreia do Sul marcando 4 gols nas oitavas, e tendo como próximo adversário a Croácia, que possui pontuações de ATT e DEF bastante

inferiores ao Brasil, há o aumento drástico da probabilidade do Brasil ser campeão da competição na Simulação 5. O modelo colocou a probabilidade de Portugal ser campeã maior que da Argentina e França, o que, assim como no caso do Brasil, pode ser resultado de enfrentar o time de Marrocos nas quartas.

Tabela 3.16: Probabilidades de cada time atingir determinada etapa da competição nas fases eliminatórias na Simulação 6 (antes dos jogos das semis).

	Probabilidades			
Time	Campeão	Vice	Terceiro	Quarto
França	0,515	0,307	0,126	0,052
Argentina	0,359	0,365	0,191	0,085
Croácia	0,076	0,2	0,376	0,348
Marrocos	0,05	0,128	0,307	0,515

Na Simulação 6, ou seja, antes dos jogos das semis, nota-se que a França tem uma probabilidade maior que a Argentina de ganhar a competição, Croácia está com maior chance de terminar em terceiro e Marrocos em quarto.

Tabela 3.17: Probabilidades de cada time atingir determinada etapa da competição nas fases eliminatórias na Simulação 7 (antes dos jogos da final e disputa de 3º lugar).

Time	Probabilidade de campeão
França	0,51
Argentina	0,49

Time	Probabilidade 3º lugar
Croácia	0,57
Marrocos	0,43

Após o modelo "aprender" com toda a competição, é chegada a hora da simulação 7, que tenta predizer a final e disputa de terceiro lugar. O modelo preveu que ambas as partidas seriam bastante equilibradas, particularmente no caso da final. Este comportamento pôde ser evidenciado em ambos os jogos, em que a Argentina venceu na disputa de pênaltis, e a Croácia venceu Marrocos por 2 x 1 com estatísticas bastante similares, como posse de bola (51% x 49%), chutes (12 x 9), faltas (13 x 11) e passes (487 x 472).

Capítulo 4

Conclusões

Este trabalho teve como objetivo prever resultados das partidas e da trajetória do torneio da Copa do Mundo de 2022. Como a competição ocorre apenas a cada quatro anos, a base dados históricos é escassa e pode não representar o potencial atual de cada time. Porém, seus jogadores no geral atuam em outras partidas regularmente, principalmente por seus clubes, mas também em amistosos e outras competições por seus países. Em outras palavras, é possível ter uma base de dados maior e mais atualizada a partir do desempenho desses jogadores nessas outras competições.

Todos os anos, uma versão do jogo eletrônico de simulação de futebol FIFA é lançado, e ele calcula a força de cada time a partir da atribuição de diversas pontuações a cada jogador em diferentes áreas e habilidades. Estas pontuações tentam justamente representar a performance atual de cada jogador a partir de seu histórico recente de competições. O modelo, a partir do jogo FIFA 23 (lançado um mês antes da competição) utilizou as as pontuações gerais (*overall*) de cada jogador para atribuir forças de ATT e DEF para cada time com participação na competição. Estas forças são usadas para confrontar os times em cada rodada. As informações de ATT e DEF de cada confronto são então armazenadas e usadas para atualizar os parâmetros do modelo preditivo Bayesiano, juntamente com *odds* de sites de aposta esportiva, substituídas a cada rodada pelas novas expectativas dos sites de aposta esportiva, além do número de gols de cada time, este utilizado de forma cumulativa ao longo das rodadas (assim como ATT e DEF).

Com estas informações, foi possível calcular probabilidades de vitória, empate e derrota em um modelo que "aprende" continuamente ao longo da competição.

Na Fase de Grupos, a incorporação das atualizações no modelo rodada a rodada inicialmente aumentou a imprecisão das previsões da rodada 1 para a 2, e manteve o mesmo

grau da rodada 2 para a 3.

Este modelo atribuiu ao Brasil a maior probabilidade de ser campeão da competição até sua queda nas quartas de final para a Croácia. Argentina e França, campeã e vice-campeã, respectivamente, firmaram-se como fortes candidatas ao título a partir das oitavas de final, chegando a final com praticamente a mesma chance de conquistar o título.

A Croácia foi o time que o modelo apresentou maior dificuldade de previsão, pois o time tinha a menor probabilidade de atingir as semi-finais entre todos os times classificados para as fases eliminatórias, dado que possuía o Brasil como adversário. Na disputa do terceiro lugar, porém, o modelo deu leve favoritismo ao time, que acabou ganhou a partida contra o Marrocos.

Uma possibilidade de aprimoramento do modelo poderia ser a partir do uso de outras distribuições de probabilidade, assim como a incorporação de outras pontuações mais específicas de cada jogador dentro jogo FIFA 23, que vão desde habilidades tradicionais como cruzamento, chutes a gol e cabeceio, até habilidades mais específicas como passes longos, reação, equilíbrio e posicionamento.

Apêndice A

Demonstrações

A.1 Posteriori Round 1

$$
\begin{aligned}
P(\lambda_A|X_A) &\propto f(\lambda_A) \times P(X_A|\lambda_A) \\
&\propto \frac{\beta_A^{\alpha_A}}{\Gamma(\alpha_A)} \lambda_A^{\alpha_A - 1} e^{-\beta_A \lambda_A} \times e^{-\mu_{A,B}} \frac{\mu_{A,B}^{x_A}}{x_A!} \\
&\propto \frac{\beta_A^{\alpha_A}}{\Gamma(\alpha_A)} \lambda_A^{\alpha_A - 1} e^{-\beta_A \lambda_A} \times e^{-\mu_{A,B}} \frac{\left(\lambda_A \frac{ATT_A}{DEF_B}\right)^{x_A}}{x_A!} \\
&\propto e^{-(\frac{ATT_A}{DEF_B} + \beta_A)\lambda_A} \lambda_A^{x_A + \alpha_A - 1} \rightarrow \\
&\rightarrow \therefore \boxed{\lambda_A|X_A = x_A \sim Gama\left(x_A + \alpha_A, \frac{ATT_A}{DEF_B} + \beta_A\right)}
\end{aligned}
$$

A.2 Preditiva Round 1

$$
\begin{aligned}
P(X_A = x_A) &= \int_{\lambda\in\Lambda} P(X_A \cap \lambda_A)\, d\lambda_A = \\
&= \int_0^\infty f(\lambda_A) \times P(X_A|\lambda_A)\, d\lambda_A = \\
&= \int_0^\infty \frac{\beta_A^{\alpha_A}}{\Gamma(\alpha_A)} \lambda_A^{\alpha_A-1} e^{-\beta_A\lambda_A} \times e^{-\mu_{A,B}} \frac{\mu_{A,B}^{x_A}}{x_A!}\, d\lambda_A = \\
&= \int_0^\infty \frac{\beta_A^{\alpha_A}}{\Gamma(\alpha_A)} \lambda_A^{\alpha_A-1} e^{-\beta_A\lambda_A} \times e^{-\left(\lambda_A \frac{ATT_A}{DEF_B}\right)} \frac{\left(\lambda_A \frac{ATT_A}{DEF_B}\right)^{x_A}}{x_A!}\, d\lambda_A = \\
&= \frac{{\beta_A}^{\alpha_A}}{\Gamma(\alpha_A)x_A!} \left(\frac{ATT_A}{DEF_B}\right)^{x_A} \int_0^\infty \lambda_A^{x_A+\alpha_A-1} e^{-\left(\frac{ATT_A}{DEF_B}+\beta_A\right)}\, d\lambda_A = \\
&= \frac{{\beta_A}^{\alpha_A}}{\Gamma(\alpha_A)x_A!} \left(\frac{ATT_A}{DEF_B}\right)^{x_A} \frac{\Gamma(x_A+\alpha_A)}{\left(\frac{ATT_A}{DEF_B}+\beta_A\right)^{x_a+\alpha_A}} \\
&\qquad \times \int_0^\infty \frac{\left(\frac{ATT_A}{DEF_B}+\beta_A\right)^{(x_A+\alpha_A)}}{\Gamma(x_A+\alpha_A)} \lambda_A^{x_A+\alpha_A-1} e^{-\left(\frac{ATT_A}{DEF_B}+\beta_A\right)}\, d\lambda_A = \\
&= \left(\frac{\beta_A}{\frac{ATT_A}{DEF_B}+\beta_A}\right)^{\alpha_A} \left(\frac{\frac{ATT_A}{DEF_B}}{\frac{ATT_A}{DEF_B}+\beta_A}\right)^{x_A} \binom{x_A+\alpha_A-1}{x_a} \rightarrow \\
&\rightarrow \therefore \boxed{X_A \sim BN\left(\alpha_A, \frac{\beta_A}{\frac{ATT_A}{DEF_B}+\beta_A}\right)}
\end{aligned}
$$

A.3 Posteriori com Jogos

$$
\begin{aligned}
P(\lambda_A|X_{A1},\ldots,X_{An}) &\propto f(\lambda_A) \times P(X_{A1},\ldots,X_{An}|\lambda_A) \\
&\propto \frac{\beta_A^{\alpha_A}}{\Gamma(\alpha_A)} \lambda_A^{\alpha_A-1} e^{-\beta_A\lambda_A} \times \frac{e^{(-\sum_{i=1}^n \mu_{Ai})} \mu_{Ai}^{\sum_{i=1}^n x_{Ai}}}{\prod_{i=1}^n x_{Ai}!} \propto \\
&\propto e^{-\left(\sum_{i=1}^n \frac{ATT_A}{DEF_B}+\beta_A\right)\lambda_A} \lambda_A^{\sum_{i=1}^n x_{Ai}+\alpha_A-1} \rightarrow \\
&\rightarrow \therefore \boxed{\lambda_A|X_{A1},\ldots,X_{An} \sim Gama\left(\sum_{i=1}^n x_{Ai}+\alpha_A, \sum_{i=1}^n \frac{ATT_A}{DEF_B}+\beta_A\right)}
\end{aligned}
$$

A.4 Preditiva com Jogos

Pelo Teorema de Bayes, sabe-se que

$$P(\lambda_A|X_A = x_a, \mathbf{X} = \mathbf{x}) = \frac{P(\lambda_A|\mathbf{x})P(X_A|\lambda_A, \mathbf{x})}{P(X_A|\mathbf{x})} =$$
$$= P(X_A|\mathbf{x}) = \frac{P(\lambda_A|\mathbf{x})P(X_A|\lambda_A)}{P(\lambda_A|X_A = x_a, \mathbf{X} = \mathbf{x})}.$$

Substituindo as expressões, temos

$$P(X_A|\mathbf{X} = \mathbf{x}) = \frac{\frac{\left(\frac{ATT_A}{DEF_B}+\beta_A\right)^{\sum_{i=1}^n x_{Ai}+\alpha_A}}{\Gamma(\sum_{i=1}^n x_{Ai}+\alpha_A)}\lambda_A^{\sum_{i=1}^n x_{Ai}+\alpha_A-1}e^{-\lambda_A\left(\sum_{i=1}^n \frac{ATT_A}{DEF_B}+\beta_A\right)}e^{-\left(\lambda_A\frac{ATT_A}{DEF_B}\right)}\frac{(\lambda_A\frac{ATT_A}{DEF_B})^x}{x!}}{\frac{\left(\sum_{i=1}^n \frac{ATT_A}{DEF_B}+\frac{ATT_A}{DEF_B}+\beta_A\right)^{\sum_{i=1}^n x_{Ai}+x+\alpha_A}}{\Gamma(\sum_{i=1}^n x_{Ai}+x+\alpha_A)}\lambda_A^{\sum_{i=1}^n x_{Ai}+x+\alpha-1}e^{-\lambda_A\left(\sum_{i=1}^n \frac{ATT_A}{DEF_{Bi}}+\frac{ATT_A}{DEF_B}+\beta_A\right)}} =$$

$$= \left[\frac{\sum_{i=1}^n \frac{ATT_A}{DEF_{Bi}}+\beta_A}{\sum_{i=1}^n \frac{ATT_A}{DEF_{Bi}}+\frac{ATT_A}{DEF_B}+\beta_A}\right]^{\sum_{i=1}^n x_{Ai}+\alpha_A}\left[\frac{1}{\sum_{i=1}^n \frac{ATT_A}{DEF_{Bi}}+\frac{ATT_A}{DEF_B}+\beta_A}\right]^x$$
$$\times\binom{\sum_{i=1}^n x_{Ai}+\alpha_A+x-1}{x} \rightarrow$$

$$\rightarrow \therefore \boxed{X_A|\mathbf{X} = \mathbf{x} \sim BN\left(\sum_{i=1}^n x_{Ai}+\alpha_A, \frac{\sum_{i=1}^n \frac{ATT_A}{DEF_{Bi}}+\beta_A}{\sum_{i=1}^n \frac{ATT_A}{DEF_{Bi}}+\frac{ATT_A}{DEF_B}+\beta_A}\right)}$$

A.5 Tabelas Adicionais

Tabela A.1: *Odds* usadas na Simulação 1.

Rodada	Time_A	Time_B	Odds_A	Odds_empate	Odds_B
1	QAT	ECU	3,5	3,2	2,46
1	ENG	IRN	1,39	4,8	13
1	SEN	NED	6,6	3,9	1,63
1	USA	WAL	2,72	3,05	3,05
1	ARG	KSA	1,2	7,5	21
1	DEN	TUN	1,48	4,33	8,8
1	MEX	POL	2,76	3,15	2,91
1	FRA	AUS	1,26	6,5	15
1	MAR	CRO	4,6	3,4	1,96
1	GER	JPN	1,52	4,7	7
1	ESP	CRC	1,16	8	26
1	BEL	CAN	1,49	4,7	7,5
1	SUI	CMR	1,75	3,6	6
1	URU	KOR	1,8	3,52	5,5
1	POR	GHA	1,44	4,5	9,5
1	BRA	SRB	1,44	4,9	8,1
2	WAL	IRN	2,33	3,2	3,61
2	QAT	SEN	4,2	3,3	2,1
2	NED	ECU	1,61	4,2	6
2	ENG	USA	1,65	4,1	6
2	TUN	AUS	2,84	3,15	2,79
2	POL	KSA	1,65	3,99	6,2
2	FRA	DEN	2,04	3,41	4,1
2	ARG	MEX	1,6	4	7
2	JPN	CRC	1,8	3,64	5,3
2	BEL	MAR	1,58	4,33	6,5
2	CRO	CAN	1,8	3,8	5,1
2	ESP	GER	2,68	3,52	2,8
2	CMR	SRB	5	3,51	1,84
2	KOR	GHA	2,76	3,1	2,97
2	BRA	SUI	1,52	4,4	7
2	POR	URU	2,15	3,39	4
3	NED	QAT	1,33	6	11
3	ECU	SEN	2,87	3,36	2,8
3	IRN	USA	3,96	3,41	2,08
3	WAL	ENG	6,5	4,1	1,61
3	TUN	FRA	11	5,4	1,35
3	AUS	DEN	7,2	4,2	1,58
3	POL	ARG	5,6	4	1,7
3	KSA	MEX	6	4	1,7
3	CRO	BEL	3,53	3,5	2,25
3	CAN	MAR	2,3	3,45	2,63
3	JPN	ESP	8	4	1,5
3	CRC	GER	16	6,25	1,28
3	KOR	POR	7,1	4,3	1,54
3	GHA	URU	5,1	3,6	1,83
3	SRB	SUI	2,87	3,3	2,75
3	CMR	BRA	10	6,2	1,33

Tabela A.2: *Odds* usadas na Simulação 2.

Rodada	Time_A	Time_B	Odds_A	Odds_empate	Odds_B
2	WAL	IRN	2,22	3,2	3,89
2	QAT	SEN	6,4	3,75	1,67
2	NED	ECU	1,8	3,75	5,4
2	ENG	USA	1,53	4,6	7,1
2	TUN	AUS	2,2	3,35	3,85
2	POL	KSA	1,78	3,78	5,3
2	FRA	DEN	1,8	3,75	5,3
2	ARG	MEX	1,6	4,1	6,8
2	JPN	CRC	1,46	4,6	8,6
2	BEL	MAR	2,05	3,6	4,1
2	CRO	CAN	2,15	3,47	3,75
2	ESP	GER	2,4	3,75	3,03
2	CMR	SRB	5,5	3,84	1,73
2	KOR	GHA	2,71	3,15	3
2	BRA	SUI	1,43	5	9
2	POR	URU	2	3,53	4,33
3	NED	QAT	1,19	8,5	23
3	ECU	SEN	2,9	3,3	2,72
3	IRN	USA	4,8	3,8	1,84
3	WAL	ENG	7	4	1,59
3	TUN	FRA	11	5	1,4
3	AUS	DEN	8,5	4,9	1,46
3	POL	ARG	7,9	4,5	1,5
3	KSA	MEX	6	4,1	1,68
3	CRO	BEL	3,11	3,47	2,42
3	CAN	MAR	3	3,45	2,55
3	JPN	ESP	11	5,6	1,3
3	CRC	GER	34	12	1,11
3	KOR	POR	6,4	4,4	1,6
3	GHA	URU	5,2	3,6	1,83
3	SRB	SUI	2,65	3,38	2,84
3	CMR	BRA	19	7,4	1,22

Tabela A.3: *Odds* usadas na Simulação 3.

Rodada	Time_A	Time_B	Odds_A	Odds_empate	Odds_B
3	NED	QAT	1,24	7	17
3	ECU	SEN	2,5	3,25	3,25
3	IRN	USA	4,1	3,55	2,06
3	WAL	ENG	8,8	4,5	1,48
3	TUN	FRA	8	4,3	1,5
3	AUS	DEN	7,3	4,4	1,54
3	POL	ARG	8	4,3	1,5
3	KSA	MEX	4,9	4,4	1,72
3	CRO	BEL	2,74	3,45	2,74
3	CAN	MAR	3,6	3,47	2,19
3	JPN	ESP	9	4,8	1,45
3	CRC	GER	26	11	1,12
3	KOR	POR	5,75	4,33	1,67
3	GHA	URU	5,4	4	1,74
3	SRB	SUI	2,56	3,4	2,99
3	CMR	BRA	12,5	6,25	1,28

Tabela A.4: *Odds* usadas na Simulação 4.

Rodada	Time_A	Time_B	Odds_A	Odds_empate	Odds_B
4	NED	USA	2,04	3,35	4,4
4	ARG	AUS	1,23	7	17
4	JPN	CRO	4	3,26	2,2
4	BRA	KOR	1,29	6,3	15
4	ENG	SEN	1,57	4	8
4	FRA	POL	1,34	5,5	13
4	MAR	ESP	7	4	1,62
4	POR	SUI	1,94	3,55	4,6

Tabela A.5: *Odds* usadas na Simulação 5.

Rodada	Time_A	Time_B	Odds_A	Odds_empate	Odds_B
5	NED	ARG	3,7	3,21	2,26
5	CRO	BRA	10	5,2	1,38
5	ENG	FRA	3,13	3,33	2,55
5	MAR	POR	6	3,79	1,71

Tabela A.6: *Odds* usadas na Simulação 6.

Rodada	Time_A	Time_B	Odds_A	Odds_empate	Odds_B
6	ARG	CRO	1,94	3,4	5,1
6	FRA	MAR	1,56	4	8

Tabela A.7: *Odds* usadas na Simulação 7.

Rodada	Time_A	Time_B	Odds_A	Odds_empate	Odds_B
6	CRO	MAR	2,4	3,5	3,2
6	ARG	FRA	2,82	3,14	2,9

A.6 Figuras Adicionais

Figura A.1: Probabilidades de cada time terminar em determinada posição dentro do grupo após a Simulação 1 (grupos A,B,C,D).

Figura A.2: Probabilidades de cada time terminar em determinada posição dentro do grupo após a Simulação 1 (grupos E,F,H).

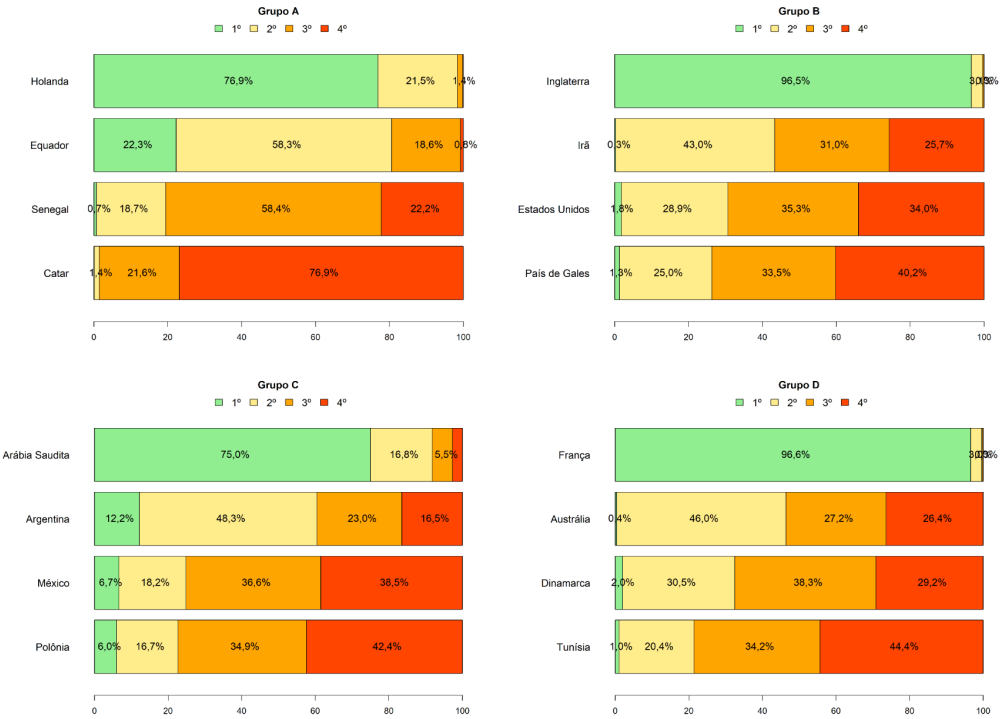

Figura A.3: Probabilidades de cada time terminar em determinada posição dentro do grupo após a Simulação 2 (grupos A,B,C,D).

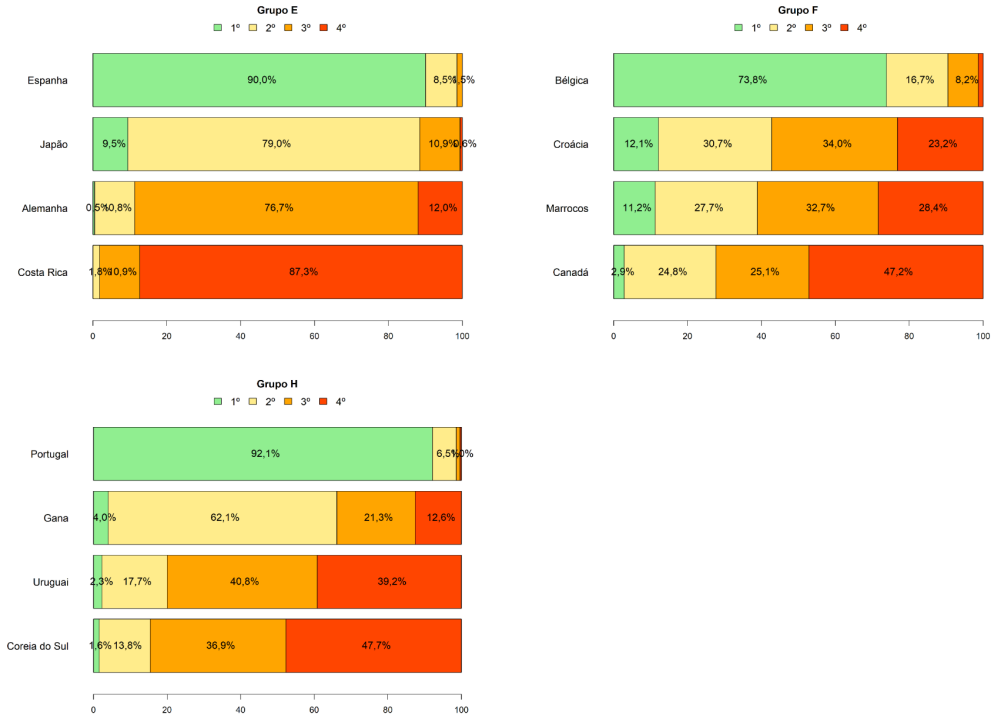

Figura A.4: Probabilidades de cada time terminar em determinada posição dentro do grupo após a Simulação 2 (grupos E,F,H).

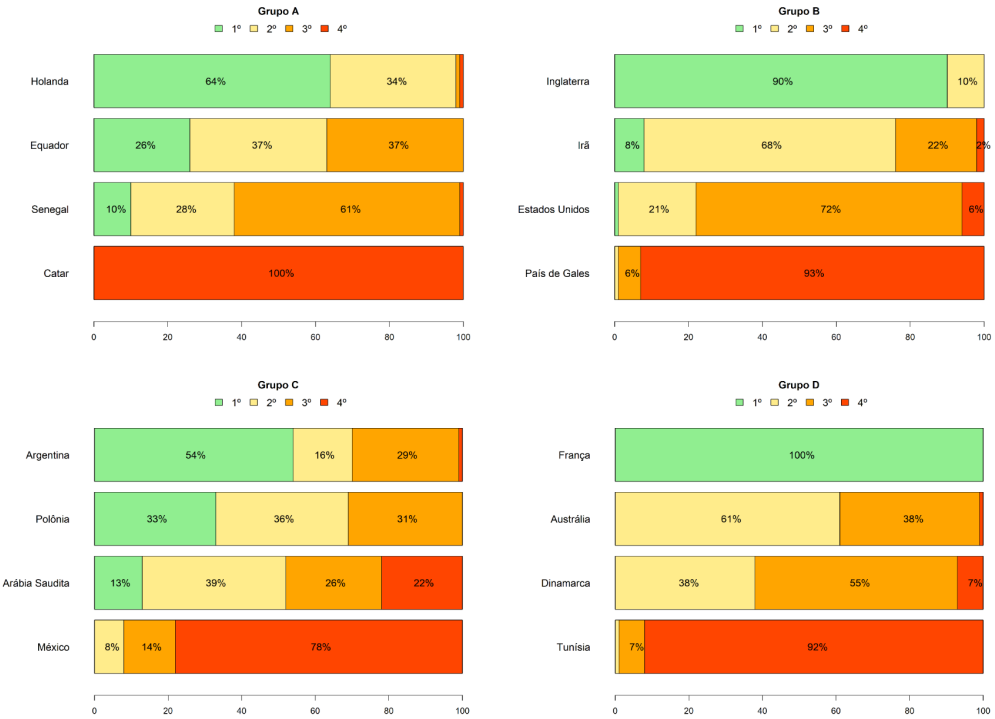

Figura A.5: Probabilidades de cada time terminar em determinada posição dentro do grupo após a Simulação 3 (grupos A,B,C,D).

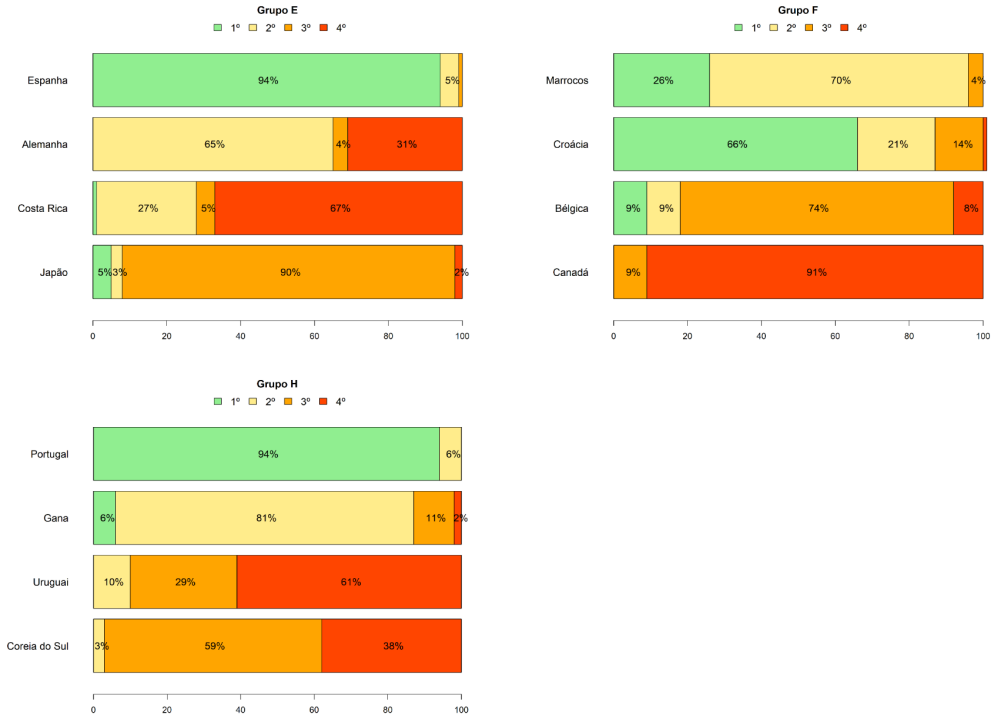

Figura A.6: Probabilidades de cada time terminar em determinada posição dentro do grupo após a Simulação 3 (grupos E,F,H).

Bibliografia

Dyte, D. e Clarke, S. R. (2000). A ratings based poisson model for world cup soccer simulation. *Journal of the Operational Research society*, **51**(8), 993–998.

Fletcher, R. (1987). *Practical Methods of Optimization*. John Wiley & Sons, New York, NY, USA, second edition.

Groll, A., Ley, C., Schauberger, G. e Van Eetvelde, H. (2019). A hybrid random forest to predict soccer matches in international tournaments. *Journal of quantitative analysis in sports*, **15**(4), 271–287.

Lee, A. J. (1997). Modeling scores in the premier league: is manchester united really the best? *Chance*, **10**(1), 15–19.

Liu, D. C. e Nocedal, J. (1989). On the limited memory bfgs method for large scale optimization. *Mathematical programming*, **45**(1-3), 503–528.

Suzuki, A. K., Salasar, L. E. B., Leite, J. e Louzada-Neto, F. (2010). A bayesian approach for predicting match outcomes: the 2006 (association) football world cup. *Journal of the Operational Research Society*, **61**(10), 1530–1539.