

MODELAGEM MATEMÁTICA PARA PREVISÃO DE JOGOS DE FUTEBOL

Luiz Fernando Pinheiro Ramos

Centro Federal de Educação Tecnológica de Minas Gerais - CEFET-MG

luizramos@cefetmg.br

Humberto Cesar Fernandes Lemos

Universidade Federal de São João del-Rei - UFSJ

humbertolemos@ufsj.edu.br

Ben Dêivide de Oliveira Batista

Universidade Federal de São João del-Rei - UFSJ

ben.deivide@ufsj.edu.br

Resumo

Este artigo se propõe a construir um método para calcular probabilidades no futebol e comparar os resultados com o modelo do Departamento de Matemática da UFMG. Na metodologia apresentada, assumimos que os números de gols marcados pelos times em uma partida são independentes e seguem uma distribuição Poisson, em que a média reflete o fator ataque ou fator defesa. Aplicada a 250 jogos da Série A do Campeonato Brasileiro de 2018, a metodologia estudada apresentou resultados coerentes com o modelo da UFMG. Foi possível prever com boa precisão o campeão e obteve resultados satisfatórios para os classificados para a Copa Libertadores da América, como também para os times rebaixados. A média da medida de DeFinetti foi 0,568 e aproximadamente 57% dos resultados estão abaixo de $2/3$. A metodologia proposta foi implementada no software R.

Abstract

Abstract This study proposes to construct a method to calculate soccer probabilities and compare the results with the UFMG Mathematics Department model. In the presented methodology, we assume that the number of goals scored by the teams in a match are independent and follow a Poisson distribution, where the average reflects the attack factor or defense factor. Applied in 250 games of Series A of the 2018 Brazilian Championship, the methodology studied presented results consistent with the UFMG model. It was possible to predict with good accuracy the champion and obtained satisfactory results for the qualifiers for the Copa Libertadores of America, as well as for the relegated teams. The DeFinetti measurement average was 0.568 and approximately 57 % of the results are below $2/3$. The proposed methodology was implemented in the software R.

1 Introdução

Uma das maiores paixões dos brasileiros é, sem dúvida, o futebol. Este esporte emocionante e uns dos assuntos mais frequentes nos meios de comunicação. Para [4], o futebol é tão encantador pela sua magia e dinamismo, que até aqueles que não são fãs declarados param para ver a seleção de seu país disputar uma Copa do Mundo.

O futebol se torna cada vez mais alvo de especialistas que buscam na matemática, através da probabilidade, meios para explicar e prever os resultados. Tal fato é tão impressionante que muitos especialistas da área esportiva dizem que o futebol é o esporte no qual a chance do mais fraco vencer o mais forte se maximiza em relação a qualquer outro esporte. Quantas vezes clubes que fizeram grandes investimentos foram derrotados por outros que, “em teoria”, eram mais fracos? Um elenco de ponta e um esquema tático infalível são primordiais nesse jogo, que às vezes parece um combate estratégico duelado pelos técnicos, mas nem sempre os melhores times vencem. É exatamente isso que torna o futebol tão impressionante [4].

Diante disso, o futebol sempre foi alvo de especulações sobre quem seria o campeão, o vice-campeão, quais os clubes seriam rebaixados, qual o time que marcaria mais gols, entre muitas outras perguntas relacionadas. Surgiram então os primeiros modelos estatísticos para tentar responder essas perguntas que muitos torcedores e admiradores do esporte faziam [6].

Em [10], o autor verificou por meio da porcentagem de vitórias obtidas como mandante que um time tem vantagem ao jogar em casa com o apoio da torcida. Em [11], propuseram uma metodologia bayesiana para calcular probabilidades das partidas de futebol, utilizando opiniões de especialistas e o ranking da FIFA como informações *a priori*. Também é possível encontrar trabalhos que utilizam a distribuição de Poisson: em [8] o autor utiliza esta distribuição para modelar os números de gols marcados pela Inglaterra, Irlanda, Escócia e País de Gales, no Campeonato Internacional Britânico 1883 à 1980. Em [9], o autor também considera a distribuição de Poisson em que as médias dependem dos efeitos ofensivos e defensivos de cada equipe.

Este trabalho propõe uma metodologia, baseada na distribuição de Poisson, para calcular probabilidades no futebol. Particularmente, usamos as distribuições de Poisson para tentar prever os gols feitos e sofridos por cada equipe do Campeonato Brasileiro de 2018, quando jogam como mandante ou visitante, para tentar prever os resultados das partidas e a classificação final de cada uma no certame. Na seção 2, detalharemos melhor o método proposto por este trabalho, o qual será chamado de método do PROFMAT, e na seção 3 confrontaremos os resultados com o modelo do Departamento de Matemática da UFMG (modelo da UFMG), que é um modelo que corrobora com o presente trabalho, pois é de referência e respeitado nacionalmente.

2 Método Proposto

Em [2], os autores do modelo UFMG argumentam que o modo de se estimar probabilidades em futebol é simples e de certa forma comum aos diferentes modelos. A ideia é partir da situação atual do campeonato e simular os jogos restantes. Ao final, o computador gera a classificação dos times e a registra. Depois, volta a fazer o mesmo inúmeras vezes, e ao término do processo já está apto a divulgar as probabilidades que nos interessam. Se, por exemplo, quisermos saber a probabilidade de um dado time ser campeão, basta ver a proporção entre as simulações em que esta equipe terminou líder sobre o total de simulações.

Para realizar esta tarefa, nosso modelo (PROFMAT) atribui, a cada time participante do campeonato, dois vetores de médias

$$\text{Mandante} = (GFM; GSM) \quad e \quad \text{Visitante} = (GFV; GSV), \quad (2.1)$$

onde GFM é a média de gols feitos como mandante, GSM é a média de gols sofridos como mandante, GFV é a média de gols feitos como visitante, GSV é a média de gols sofridos como visitante. Para [2], a ideia de se tratar um mesmo time como dois diferentes – um em casa, e outro fora dela – não é nenhum tipo de esquizofrenia futebolística. Ao contrário, reflete apenas o modo como a torcida, ou o próprio campo onde a equipe habitualmente joga, podem influenciar no desempenho do clube. Em suma, jogos iguais em campos distintos, são distintos.

De acordo com [2], a questão chave é como atualizar os vetores de médias de duas equipes que se enfrentam, após a realização do jogo. A premissa da qual partimos é a de que, na primeira rodada do campeonato, quando $r = 0$, todos os times recebem as mesmas médias GFM, GSM, GFV e GSV, que são as médias globais do último campeonato. Em 2017, por exemplo, essas médias foram $GFM = GSV = 1,38$ e $GFV = GSM = 1,04$, então na primeira rodada do campeonato de 2018, todos os times recebem esses dois vetores de médias

$$\text{Mandante} = (1,38; 1,04) \quad e \quad \text{Visitante} = (1,04; 1,38). \quad (2.2)$$

A cada rodada, os vetores são realimentados. Se o time jogou como mandante, muda-se o seu vetor mandante, e o mesmo vale para seu vetor visitante, caso o jogo tenha sido fora de casa. Essas médias são atualizadas ao longo do campeonato da seguinte maneira:

$$GFM = \frac{\sum_{i=0}^r (\text{gols feitos como mandante})}{r + 1}, \quad (2.3)$$

$$GSM = \frac{\sum_{i=0}^r (\text{gols sofridos como mandante})}{r+1}, \quad (2.4)$$

$$GFV = \frac{\sum_{i=0}^r (\text{gols feitos como visitante})}{r+1}, \quad (2.5)$$

$$GSV = \frac{\sum_{i=0}^r (\text{gols sofridos como visitante})}{r+1}, \quad (2.6)$$

onde r é quantidade de rodadas do campeonato já realizadas e i é o índice que percorre essas rodadas – veja que o somatório inicia-se em zero, pois na primeira rodada do campeonato não temos nenhuma rodada realizada, e usamos os vetores (2.2) como valores iniciais.

Por exemplo, na primeira rodada do campeonato brasileiro de 2018, o Grêmio venceu o Cruzeiro por 1 a 0 no Mineirão. Então para simular os jogos destas equipes nas rodadas seguintes do campeonato, os vetores de médias dessas equipes são atualizados da seguinte forma

$$CRU_{mandante} = \left(\frac{1.38 + 0}{2}; \frac{1.04 + 1}{2} \right) \quad e \quad CRU_{visitante} = (1.38; 1.04). \quad (2.7)$$

$$GRE_{mandante} = (1.38; 1.04) \quad e \quad GRE_{visitante} = \left(\frac{1.04 + 1}{2}; \frac{1.38 + 0}{2} \right). \quad (2.8)$$

Como o Cruzeiro jogou em casa e o Grêmio fora, os vetores $CRU_{visitante}$ e $GRE_{mandante}$ não são alterados. Perceba que estamos dividindo por 2, pois temos apenas uma rodada do campeonato realizada ($r = 1$), ou seja, estamos calculando a média entre os resultados da primeira rodada com o parâmetro do campeonato passado, se por exemplo, estivermos simulando a 21ª rodada (quando $r = 20$), a soma do parâmetro do campeonato passado com os resultados da 20 rodadas passadas será dividida por 21. Seguindo esse raciocínio, a partir de uma rodada qualquer, simulamos todo o campeonato e armazenamos os resultados.

Com as médias GFM, GSM, GFV e GSV bem definidas, para simular o resultado de uma única partida criamos o fator mandante $(GFM + GSV)/2$ e o fator visitante $(GFV + GSM)/2$. Em outras palavras, o fator mandante segue a distribuição de Poisson com parâmetro $(GFM + GSV)/2$, e o fator visitante segue a Poisson com parâmetro $(GFV + GSM)/2$. Com esses dois fatores podemos formar o vetor média do jogo que é

$$Vetor\ Jogo = \left(\frac{GFM + GSV}{2}; \frac{GFV + GSM}{2} \right), \quad (2.9)$$

em que a primeira coordenada é o fator mandante e a segunda o fator visitante.

Para exemplificar, suponha uma partida de futebol entre Atlético - MG e Chapecoense, válida pela nona rodada do campeonato. Usando os resultados reais das oito primeiras rodadas disputadas por estas equipes, temos que o fator mandante é Poisson(2,20) e que o fator visitante é Poisson(1,50), então o vetor jogo será

$$V_{CAM \times CHA} = (2, 20; 1, 50). \quad (2.10)$$

Com este vetor o computador calcula as probabilidades através da função de probabilidade da Poisson e divide o intervalo $[0, 1]$ em seis partes de acordo com as probabilidades acumuladas. As Tabelas 1 e 2 mostram as probabilidades e suas acumuladas.

Tabela 1: Probabilidade de gols para o fator mandante.

Gols	Probabilidade	Acumulada
0	0,111	0,111
1	0,244	0,355
2	0,268	0,623
3	0,197	0,819
4	0,108	0,928
5 ou mais	0,072	1,000

Fonte: Próprio autor.

Tabela 2: Probabilidade de gols para o fator visitante.

Gols	Probabilidade	Acumulada
0	0,223	0,223
1	0,335	0,558
2	0,251	0,809
3	0,126	0,935
4	0,047	0,982
5 ou mais	0,018	1,000

Fonte: Próprio autor.

A tabela 1 mostra que iremos dividir o intervalo real $[0, 1]$ da seguinte forma: $[0, 0,111]$, $(0,111, 0,355]$, $(0,355, 0,623]$, $(0,623, 0,819]$, $(0,819, 0,928]$ e $(0,928, 1]$. Para o fator visitante, a tabela 2 nos mostra que as seis partes do intervalo são: $[0, 0,223]$, $(0,223, 0,558]$, $(0,558, 0,809]$, $(0,809, 0,935]$, $(0,935, 0,982]$ e $(0,982, 1]$. Note nas tabelas que optamos por calcular as probabilidades de um time fazer até 5 gols para tornar a simulação viável, uma vez que são raros jogos em que uma das equipes faça 6 ou mais gols.

Logo após, o computador sorteia dois números aleatórios entre 0 e 1, um para o fator mandante e o outro para o fator visitante e vê em quais subintervalos os números

sorteados caíram. Se, por exemplo, o número sorteado para o fator mandante é 0,553 e para o fator visitante é 0,448, então o time mandante faz 2 gols e o time visitante faz 1 gol, assim o resultado do jogo simulado pelo computador é 2x1 para o time mandante.

Depois de simular todas as partidas de uma rodada, o programa atualiza a tabela de classificação do campeonato e os vetores médias de todos os times, de acordo com os resultados da rodada, e prossegue para a simulação das rodadas seguintes usando o mesmo procedimento descrito, até o final do torneio. Ao término da simulação, o time que ficar em primeiro lugar será contabilizado pelo programa como campeão, os seis primeiros vão para a Libertadores, do sétimo ao décimo segundo vão para sulamericana e os quatro últimos são rebaixados. Após computar todas estas ocorrências, o programa retorna à simulação da primeira rodada que ainda não foi disputada no campeonato real – no nosso exemplo da equação (2.10), retorna-se à nona rodada – e simula um novo campeonato a partir deste ponto. Assim, este processo é repetido milhares de vezes: para este trabalho, simulamos 500 mil campeonatos. A probabilidade de um time ser campeão é quantidade de vezes que ele terminou em primeiro lugar dividido pelo número total de simulações. Caso estejamos interessados nas probabilidades do resultados de um jogo, ao invés de todo um campeonato, o raciocínio é análogo: a probabilidade de vitória de um time é a quantidade de vezes que ele venceu dividido pelo número de simulações da partida.

A seguir, apresentamos um método para medir a qualidade dos resultados, as probabilidades encontradas pelo método apresentado e seu percentual de acerto comparado ao modelo do Departamento de Matemática da UFMG.

2.1 Qualidade dos Resultados

Para verificar a qualidade dos resultados do método proposto na subseção anterior, iremos utilizar a medida de DeFinetti: este método foi proposto por [7] e consiste na consideração de um plano em \mathbb{R}^3 como representação geométrica do conjunto de possíveis previsões probabilísticas. Os vértices desse plano correspondem às ocorrências dos resultados e os demais pontos a todas as outras possíveis previsões, ou seja,

$$S = \{(PVM, PEM, PVV) \in \mathbb{R}^3 : PVM + PEM + PVV = 1, \\ PVM \geq 0, PEM \geq 0, PVV \geq 0\}. \quad (2.11)$$

onde PVM denota a probabilidade de vitória do mandante, PEM a probabilidade de empate e PVV a probabilidade de vitória do visitante.

A medida de distância de DeFinetti corresponde à distância euclidiana quadrática entre o ponto correspondente à probabilidade prevista e o vértice de uma previsão

[5], um índice dado pela média aritmética das distâncias de DeFinetti, é chamado de medida DeFinetti. Para o futebol, associam-se os vértices $(1, 0, 0)$, $(0, 1, 0)$ e $(0, 0, 1)$ respectivamente à vitória da equipe mandante, ao empate, e à vitória da equipe visitante. Ao vetor de probabilidades atribuídas para uma determinada partida associa-se o ponto $(PVM, PEM, PVV) \in S$.

Assim, a distância de DeFinetti será igual a:

$$(PVM - 1)^2 + (PEM - 0)^2 + (PVV - 0)^2, \quad (2.12)$$

se a equipe mandante vencer a partida,

$$(PVM - 0)^2 + (PEM - 1)^2 + (PVV - 0)^2, \quad (2.13)$$

se a partida terminar empatada, ou

$$(PVM - 0)^2 + (PEM - 0)^2 + (PVV - 1)^2, \quad (2.14)$$

se a equipe visitante ganhar a partida.

Por exemplo, se a previsão for $(0.45, 0.30, 0.25)$ e o resultado for vitória da equipe mandante $(1, 0, 0)$, então a distância de DeFinetti é $(0.45-1)^2+(0.30-0)^2+(0.25-0)^2 = 0.3025$. Nas previsões temos que um padrão comumente utilizado é a atribuição equi-provável de probabilidades $(PVM = PEM = PVV = 1/3)$, ou seja, atribuir chances iguais a cada resultado em cada jogo. Para essa atribuição a medida de DeFinetti é igual $(1/3 - 1)^2 + (1/3 - 0)^2 + (1/3 - 0)^2 = 2/3$.

Assim, podemos considerar métodos de previsões de qualidade minimamente aceitável, aqueles que apresentam medidas de DeFinetti menores que $2/3$ e de má qualidade, aqueles que apresentam medidas maiores que $2/3$ [3].

3 Resultados

Nesta seção, apresentamos os resultados obtidos da aplicação do método no Campeonato Brasileiro de 2018. Primeiramente, realizamos uma breve análise dos resultados dos jogos. Em seguida, apresentamos os resultados para as probabilidades dos jogos de uma determinada rodada e para o campeonato inteiro. Se quisermos saber as probabilidades do campeonato para uma determinada rodada, utilizamos os resultados obtidos até a rodada anterior. Por exemplo, para as probabilidades da rodada 28, foram utilizados os resultados do campeonato até a rodada 27.

3.1 Análise Descritiva

Realizamos uma análise descritiva dos dados do Campeonato Brasileiro de 2018. A Tabela 3 indica o número de gols marcados que cada equipe fez como mandante, visitante e no geral. Como podemos ver, o Atlético-PR teve o melhor ataque como mandante (44 gols), o Flamengo teve o melhor ataque como visitante (29 gols) e o Palmeiras foi o melhor ataque do campeonato (64 gols). O Brasileirão de 2018 teve um total de 827 gols em 380 jogos, uma média de 2,18 gols por jogo. Desses 380 jogos, 202 (53%) terminaram com a vitória do mandante, 68 (18%) com a vitória do visitante e 110 (29%) com empate. O resultado mais comum foi a vitória do mandante por 1x0 (68 jogos), 238 jogos (63%) tiveram menos de três gols e 142 (37%) tiveram três ou mais gols.

Tabela 3: Número de gols marcados pelos times no campeonato.

Times	1º Turno		2º Turno		Geral	
	Mandante	Visitante	Mandante	Visitante	Mandante	Visitante
América-MG	13	5	9	3	22	8
Atlético-MG	19	14	12	11	31	25
Atlético-PR	16	2	28	8	44	10
Bahia	17	4	10	8	27	12
Botafogo	9	9	13	7	22	16
Ceará	6	4	12	10	18	14
Chapecoense	13	6	11	4	24	10
Corinthians	10	12	9	3	19	15
Cruzeiro	10	4	15	5	25	9
Flamengo	17	12	13	17	30	29
Fluminense	9	10	8	5	17	15
Grêmio	16	6	20	6	36	12
Internacional	14	12	18	7	32	19
Palmeiras	17	9	25	13	42	22
Paraná	7	2	6	3	13	5
Santos	14	4	14	14	28	18
São Paulo	15	15	10	6	25	21
Sport	11	8	8	8	19	16
Vasco da Gama	16	6	13	6	29	12
Vitória	11	9	11	5	22	14

Fonte: Próprio autor.

3.2 Probabilidades para uma partida

Ao todo, foram simulados 250 jogos e as probabilidades de vitória, empate e derrota foram calculadas da 14^a até a 38^a rodada. No método do PROFMAT, a medida

DeFinetti associadas às probabilidades desses jogos obteve uma média igual a 0,5680, enquanto que no modelo da UFMG a média foi de 0,6273. É importante ressaltar que em ambos os modelos os times começam com as mesmas chances de sucesso, e o desempenho deles ao longo do campeonato é que vai atualizando suas próprias probabilidades. Em outras palavras, os modelos aprendem com o campeonato quais times têm maiores chances de disputar o título ou de serem rebaixados, por exemplo. Assim sendo, são necessárias algumas rodadas para que eles comecem a dar previsões consistentes, pois no começo do campeonato os times ainda têm pouca história para contar. No caso do modelo PROFMAT, na primeira rodada a estimativa do parâmetro do modelo é fixado e baseado no resultado do campeonato do ano anterior. Posteriormente, a cada rodada há uma atualização da estimativa do parâmetro, e por consequência, ao longo do tempo esse valor irá convergindo para o valor real estimado. Como consequência, à medida que ocorre as rodadas, o modelo melhora na tomada de decisão.

Por exemplo, na Tabela 4 apresentamos as probabilidades para cada uma das partidas da rodada 24 juntamente com o placar observado e a medida de DeFinetti. Como já era esperado, as probabilidades de vitória são maiores quando o time é mandante.

Observe que o método do PROFMAT estima as probabilidades de vitória do mandante maiores que as estimadas no modelo da UFMG. Com relação ao empate, as probabilidades estimadas pelo modelo da UFMG são maiores que as do método do PROFMAT, fato que pode ser observado na Figura 1.

Tabela 4: Probabilidades e medida de DeFinetti para a rodada 24.

Jogo	Mandante	Visitante	Probabilidades						Placar	DeFinetti	
			Vitória		Empate		Derrota			1	2
			1	2	1	2	1	2		1	2
Jogo 1	Sport	Cruzeiro	0,398	0,218	0,321	0,510	0,186	0,272	0x0	0,698	0,362
Jogo 2	São Paulo	Bahia	0,624	0,462	0,244	0,395	0,132	0,143	1x0	0,218	0,466
Jogo 3	Flamengo	Chapecoense	0,653	0,487	0,233	0,341	0,115	0,172	2x0	0,188	0,409
Jogo 4	América	Ceará	0,474	0,306	0,292	0,431	0,234	0,263	0x0	0,780	0,486
Jogo 5	Fluminense	Botafogo	0,525	0,247	0,267	0,471	0,209	0,282	1x0	0,341	0,868
Jogo 6	Internacional	Grêmio	0,482	0,302	0,332	0,453	0,186	0,255	1x0	0,414	0,772
Jogo 7	Palmeiras	Corinthians	0,538	0,367	0,248	0,434	0,214	0,199	1x0	0,321	0,628
Jogo 8	Paraná	Santos	0,386	0,231	0,316	0,524	0,299	0,245	0x2	0,740	0,898
Jogo 9	Vitória	Vasco	0,470	0,455	0,282	0,309	0,248	0,236	1x0	0,423	0,448
Jogo 10	Atlético - MG	Atlético - PR	0,574	0,474	0,274	0,366	0,152	0,160	3x1	0,279	0,437

Legenda: 1 = PROFMAT - UFSJ e 2 = DEMAT - UFMG.

Fonte: Próprio autor e <http://www.mat.ufmg.br/futebol>.

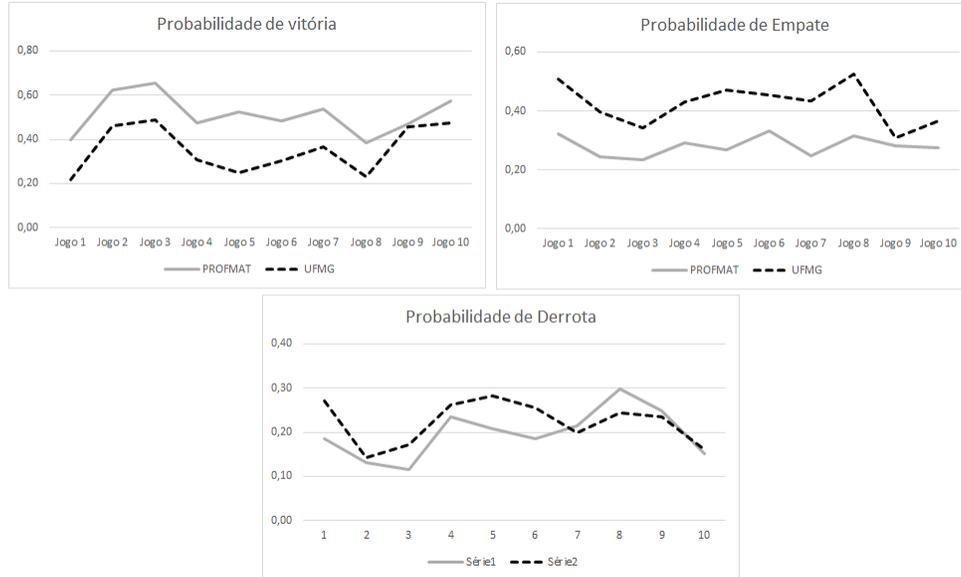


Figura 1: Probabilidades da rodada 24.

Fonte: Próprio autor e <http://www.mat.ufmg.br/futebol>.

Na Tabela 5 apresentamos o acerto de cada modelo, consideramos como acerto a ocorrência do evento com maior probabilidade. Dos 250 jogos simulados, o mandante venceu em 136, dos quais o método do PROFMAT acertou 131 e o modelo da UFMG 68. Quanto ao empate, 70 jogos terminaram com o placar igual, o método do PROFMAT não acertou nenhum resultado e o da UFMG acertou 38. Já o visitante venceu em 44 jogos, o método do PROFMAT acertou 6 e o da UFMG acertou apenas 1.

Aqui já detectamos algumas falhas do método do PROFMAT, que são a falta de acerto de empate e o pouco acerto de vitória do visitante. O ponto positivo do método é o bom acerto de vitória do mandante. Já o modelo da UFMG, acerta bem os empates e vitórias do mandantes, mas também deixa a desejar nas vitórias dos visitantes.

Na Figura 2 é possível ver a medida de DeFinetti para cada um dos jogos da rodada 24, tanto o modelo da UFMG quanto o método do PROFMAT apresentaram 3 resultados acima de $2/3$.

Tabela 5: Percentual de acerto por modelo.

Modelo	Vitória (136)	Empate (70)	Derrota (44)
PROFMAT	131 (96,32%)	0 (0,00%)	6 (13,64%)
UFMG	68 (50,00%)	38 (54,28%)	1 (2,27%)

Fonte: Próprio autor e <http://www.mat.ufmg.br/futebol>.

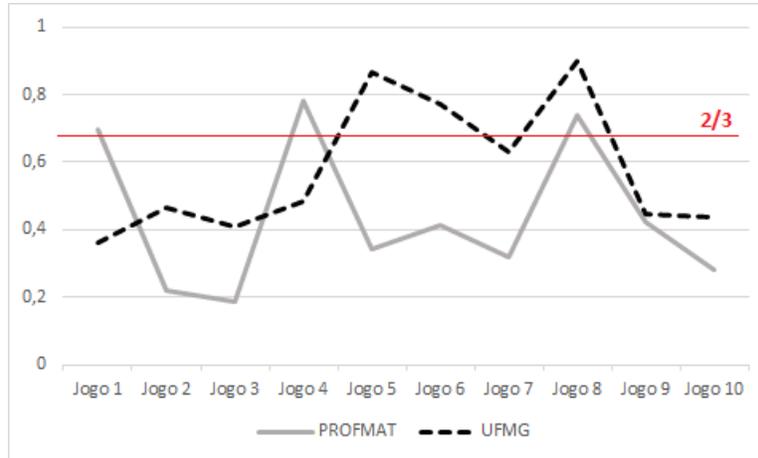


Figura 2: Medida de DeFinetti para a rodada 24.

Fonte: Próprio autor e <http://www.mat.ufmg.br/futebol>.

Na Tabela 6 apresentamos os percentis da medida de DeFinetti, no modelo da UFMG, aproximadamente 60% dos resultados estão abaixo de $2/3$, enquanto que no modelo PROFMAT esse percentual é próximo de 57%.

Tabela 6: Percentis da medida de DeFinetti.

Percentis	DeFinetti	
	PROFMAT	UFMG
0.56	0.648	0.631
0.57	0.670	0.646
0.58	0.682	0.655
0.59	0.693	0.664
0.60	0.702	0.668

Fonte: Próprio autor e <http://www.mat.ufmg.br/futebol>.

3.3 Probabilidades para a Classificação Final do Campeonato

Para obter as probabilidades de um determinado time ser campeão, de ir para Libertadores ou de ser rebaixado, realizamos 500 mil simulações para cada rodada a partir da 14ª rodada. Na Tabela 7, apresentamos as probabilidades para os seis melhores times do campeonato, tomadas ao início de cada rodada do segundo turno, de se tornarem campeões. Observe que a partir da rodada 29 o Palmeiras passou a ser o time favorito ao título, 0,569 no método do PROFMAT e 0,553 no modelo da UFMG, reflexo de uma sequência de 13 jogos sem derrota, com 3 empates e 10 vitórias.

Tabela 7: Probabilidade de conquista do título para cada um dos seis melhores times do campeonato.

Rodada	Palmeiras		Flamengo		Grêmio		Internacional		São Paulo		Atlético-MG	
	1	2	1	2	1	2	1	2	1	2	1	2
20	0,049	0,028	0,142	0,142	0,095	0,127	0,224	0,197	0,447	0,464	0,041	0,044
21	0,065	0,045	0,156	0,187	0,058	0,083	0,336	0,299	0,360	0,361	0,023	0,023
22	0,061	0,036	0,150	0,159	0,040	0,049	0,277	0,205	0,459	0,510	0,011	0,038
23	0,126	0,092	0,088	0,090	0,071	0,100	0,289	0,204	0,413	0,473	0,012	0,039
24	0,168	0,145	0,057	0,063	0,060	0,083	0,385	0,349	0,310	0,338	0,019	0,021
25	0,173	0,144	0,055	0,059	0,024	0,023	0,425	0,409	0,304	0,345	0,019	0,019
26	0,189	0,161	0,055	0,045	0,040	0,047	0,369	0,376	0,329	0,354	0,018	0,015
27	0,289	0,260	0,076	0,079	0,054	0,088	0,338	0,322	0,233	0,244	0,008	0,007
28	0,341	0,330	0,050	0,049	0,083	0,129	0,339	0,342	0,178	0,144	0,009	0,006
29	0,569	0,553	0,114	0,109	0,042	0,078	0,198	0,186	0,074	0,078	0,003	0,002
30	0,628	0,619	0,121	0,125	0,010	0,018	0,216	0,212	0,024	0,026	0,000	0,000
31	0,707	0,735	0,180	0,169	0,004	0,006	0,102	0,085	0,006	0,006	0,000	0,000
32	0,776	0,793	0,121	0,109	0,001	0,001	0,089	0,085	0,014	0,012	0,000	0,000
33	0,862	0,897	0,059	0,037	0,001	0,001	0,076	0,064	0,003	0,002	0,000	0,000
34	0,920	0,933	0,022	0,010	0,001	0,003	0,055	0,053	0,001	0,005	0,000	0,000
35	0,952	0,967	0,013	0,005	0,000	0,000	0,035	0,028	0,000	0,000	0,000	0,000
36	0,946	0,966	0,044	0,023	0,000	0,000	0,010	0,011	0,000	0,000	0,000	0,000
37	0,973	0,984	0,023	0,016	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
38	1,000	1,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000

Legenda: 1 = PROFMAT - UFSJ e 2 = DEMAT - UFMG.

Fonte: Próprio autor e <http://www.mat.ufmg.br/futebol>.

O São Paulo iniciou o segundo turno como o time com maior probabilidade de título, 0,447 no método do PROFMAT e 0,464 no modelo da UFMG, mas, nos 10 primeiros jogos do retorno, conquistou apenas 11 pontos, aproveitamento de 36,67%, com isso sua probabilidade de título foi caindo a cada rodada.

O Internacional no início do retorno era o segundo time com maior probabilidade de título, 0,224 no método do PROFMAT e 0,197 no modelo da UFMG mas, assim como o São Paulo, não teve um bom início de segundo turno, conquistando 18 pontos em 10 jogos, aproveitamento de 60%. Já o Palmeiras, nestes mesmos 10 jogos, venceu 8 e empatou 2, conquistando 26 pontos, um aproveitamento de 86,67%, fato que justifica o crescimento da probabilidade de título a cada rodada.

As probabilidades de título do método do PROFMAT estão coerentes com as do modelo da UFMG: na Figura 3, é possível ver um comparativo dessas probabilidades para cada um desses seis melhores times. Observe que de modo geral, o compartimento da curva de probabilidades do método do PROFMAT acompanha a curva do modelo da UFMG.

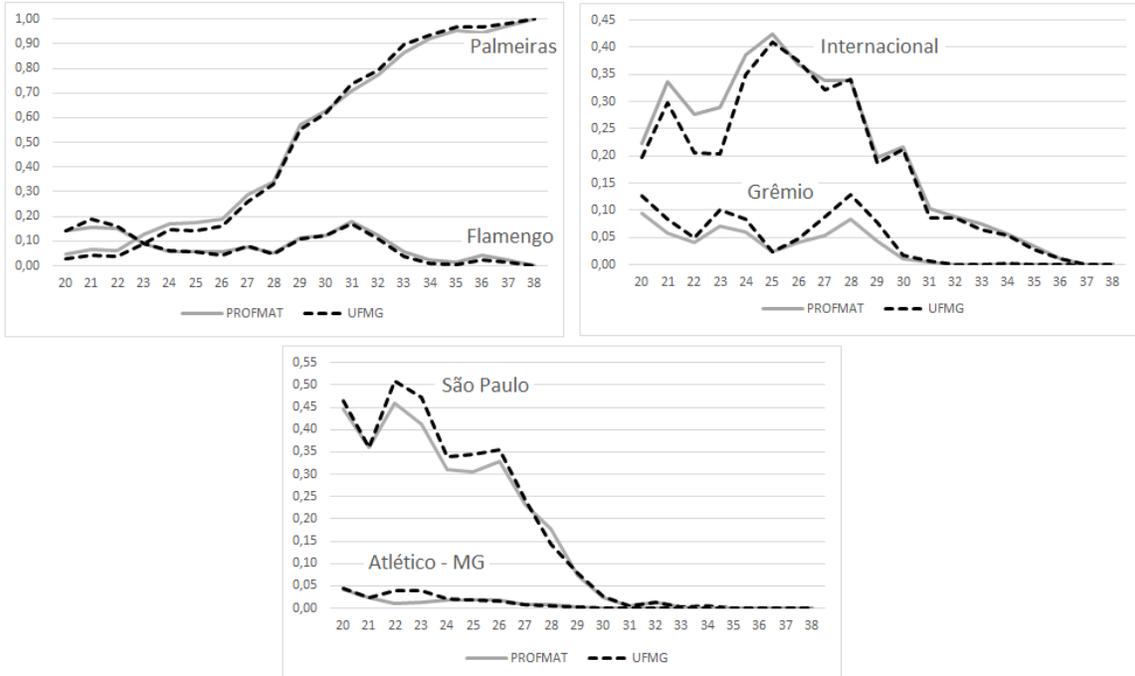


Figura 3: Probabilidade de conquista do título para cada um dos seis melhores times do campeonato.

Fonte: Próprio autor e <http://www.mat.ufmg.br/futebol>.

Na Tabela 8, apresentamos as probabilidades de conquista de uma vaga para a Libertadores de cada uma das seis melhores equipes. Todas as seis equipes iniciaram o segundo turno com probabilidade acima de 0,700. O São Paulo, que era o líder do campeonato, tinha 0,988 no método do PROFMAT e 0,989 no modelo da UFMG, mas como teve um início ruim de segundo turno, conquistou a vaga só na 32ª rodada.

O Internacional que era o vice líder, iniciou o retorno com probabilidade de conquista de uma vaga na Libertadores de 0,962 no método do PROFMAT e 0,950 no modelo da UFMG e conquistou a vaga em definitivo na 29ª rodada. Já o Palmeiras, conquistou a vaga na 28ª rodada, reflexo da sua campanha de 13 jogos sem derrota. O Flamengo conquistou a vaga na Libertadores na 30ª rodada e o Grêmio na 33ª rodada. O último time a conquistar a vaga na Libertadores foi o Atlético-MG, que entrou na última rodada disputando com o Atlético-PR, as probabilidades desta disputa podem ser vistas na Figura 4.

Tabela 8: Probabilidade de conquista de uma vaga na libertadores para cada um dos seis melhores times do campeonato.

Rodada	Palmeiras		Flamengo		Grêmio		Internacional		São Paulo		Atlético-MG	
	1	2	1	2	1	2	1	2	1	2	1	2
20	0,806	0,735	0,934	0,920	0,904	0,913	0,962	0,950	0,988	0,989	0,758	0,735
21	0,879	0,810	0,956	0,957	0,882	0,892	0,986	0,982	0,987	0,986	0,712	0,666
22	0,890	0,777	0,961	0,946	0,854	0,817	0,984	0,969	0,994	0,993	0,621	0,756
23	0,946	0,887	0,918	0,882	0,920	0,898	0,985	0,969	0,992	0,992	0,638	0,747
24	0,973	0,953	0,906	0,867	0,930	0,910	0,995	0,992	0,991	0,989	0,771	0,710
25	0,986	0,979	0,936	0,925	0,906	0,842	0,998	0,998	0,995	0,996	0,837	0,806
26	0,989	0,984	0,941	0,916	0,944	0,920	0,997	0,997	0,997	0,997	0,850	0,794
27	0,996	0,993	0,967	0,958	0,965	0,963	0,997	0,996	0,994	0,994	0,792	0,676
28	0,999	0,999	0,973	0,970	0,991	0,992	0,999	0,999	0,996	0,996	0,880	0,805
29	1,000	1,000	0,995	0,992	0,988	0,985	0,998	0,988	0,993	0,992	0,833	0,681
30	1,000	1,000	0,999	0,999	0,984	0,983	1,000	1,000	0,991	0,990	0,769	0,633
31	1,000	1,000	1,000	1,000	0,988	0,991	1,000	1,000	0,989	0,994	0,691	0,555
32	1,000	1,000	1,000	1,000	0,958	0,966	1,000	1,000	0,998	0,999	0,542	0,364
33	1,000	1,000	1,000	1,000	0,998	0,999	1,000	1,000	1,000	1,000	0,442	0,287
34	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	0,479	0,321
35	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	0,685	0,554
36	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	0,682	0,581
37	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	0,871	0,839
38	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	0,952	0,889

Legenda: 1 = PROFMAT - UFSJ e 2 = DEMAT - UFMG

Fonte: Próprio autor e <http://www.mat.ufmg.br/futebol>.

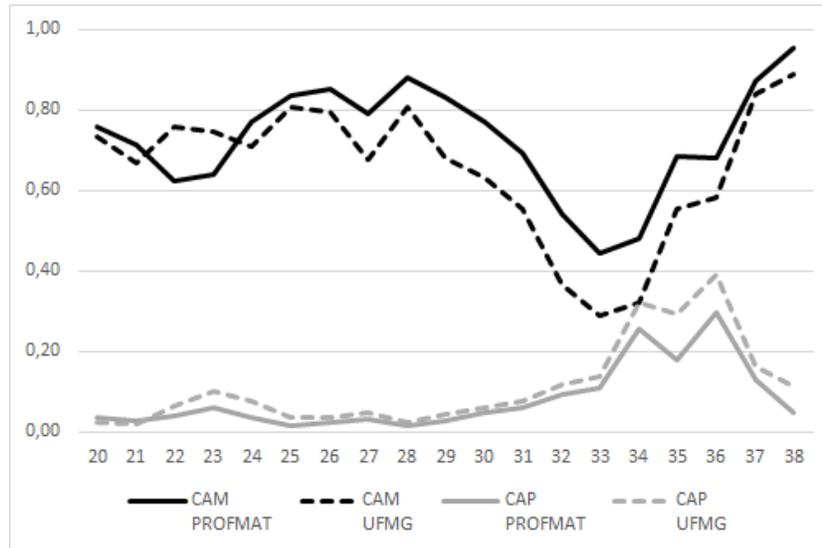


Figura 4: Probabilidade de conquista de uma vaga na Libertadores Atlético - MG (CAM) versus Atlético - PR (CAP).

Fonte: Próprio autor e <http://www.mat.ufmg.br/futebol>.

Na Figura 5, é possível ver um comparativo das probabilidades de conquista de uma vaga para a Libertadores para cada um desses seis melhores times. Observe que de modo geral, o compartimento da curva de probabilidades do método do PROFMAT acompanha a curva do modelo da UFMG.

Uma observação importante é que o Cruzeiro foi campeão da Copa do Brasil antes da 31ª rodada, mas ele não esteve entre os seis melhores e o Atlético-PR foi campeão da Sulamericana após o fim do campeonato brasileiro.

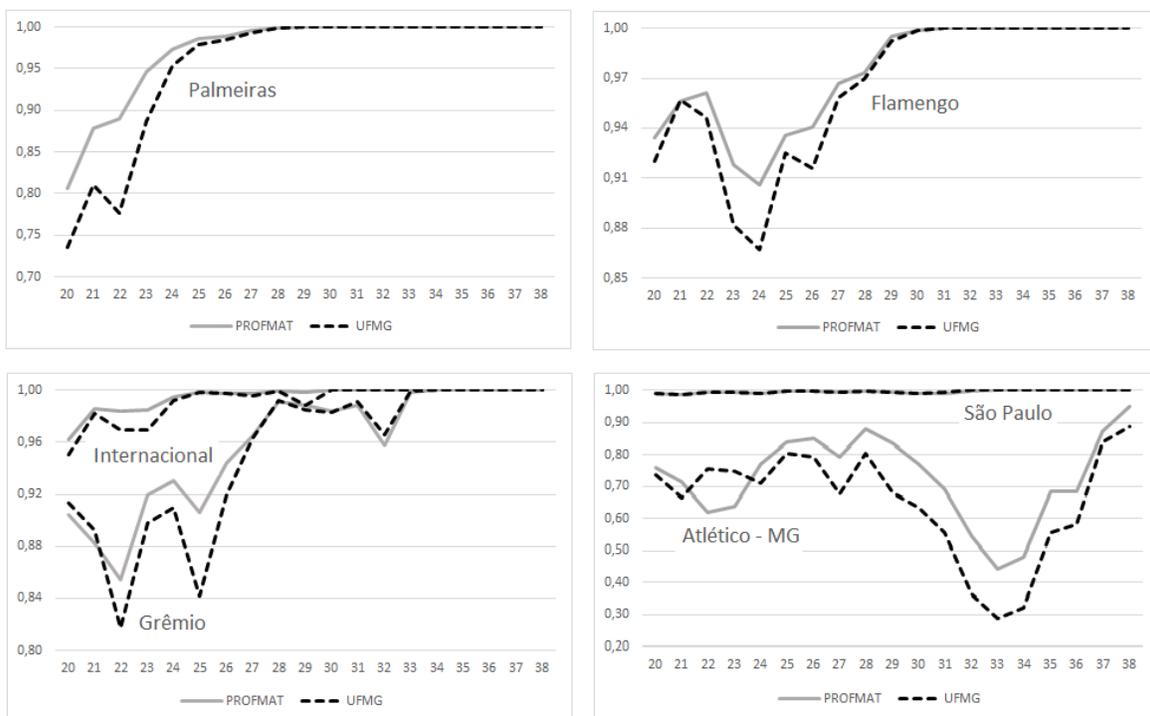


Figura 5: Probabilidade de conquista de uma vaga na Libertadores para cada um dos seis melhores times do campeonato.

Fonte: Próprio autor e <http://www.mat.ufmg.br/futebol>.

Na Tabela 9, apresentamos as probabilidades de rebaixamento das seis piores equipes, ao fim do campeonato os quatro últimos colocados caem para a segunda divisão. O Paraná, que era o lanterna do campeonato com 14 pontos ganhos e aproveitamento de 24,6%, iniciou o retorno com probabilidade de rebaixamento de 0,874 no método do PROFMAT e 0,807 no modelo da UFMG e foi matematicamente rebaixado na 29ª rodada.

O Ceará, vice lanterna com 16 pontos ganhos e aproveitamento de 28,1%, foi o time que apresentou maior discrepância nas probabilidades, iniciou o segundo turno com probabilidade de rebaixamento de 0,723 no método do PROFMAT e 0,491 no modelo da UFMG, mas venceu 5 dos 10 primeiros jogos do retorno, (inclusive venceu o Flamengo no Maracanã e Cruzeiro no Mineirão), empatou 3 e perdeu 2, aproveitamento de 60% e conquistou 18 pontos. Com esses resultados o Ceará saiu da zona de rebaixamento na 26ª rodada e foi até o fim do campeonato sem voltar para ela.

O América iniciou o retorno com probabilidade de rebaixamento de 0,223 no método do PROFMAT e 0,216 no modelo da UFMG, ele estava na 12ª colocação com 22 pontos, mas nas últimas 10 rodadas do campeonato perdeu 6 jogos, inclusive perdeu em casa para o Paraná que já estava rebaixado, e acabou rebaixado.

O vitória foi rebaixado na 37ª rodada e cinco times entraram na última rodada com chances de serem rebaixados, Vasco, Chapecoense, Fluminense, América e Sport, os dois últimos foram rebaixados.

Na Figura 6, é possível ver a probabilidades de rebaixamento para as seis piores equipes do campeonato e novamente o comportamento da curva de probabilidades do método do PROFMAT acompanha a curva do modelo da UFMG.

Tabela 9: Probabilidade de rebaixamento para os seis piores times do campeonato.

Rodada	Paraná		Vitória		América		Sport		Vasco		Ceará	
	1	2	1	2	1	2	1	2	1	2	1	2
20	0,874	0,807	0,510	0,452	0,223	0,216	0,334	0,371	0,261	0,232	0,723	0,491
21	0,879	0,800	0,530	0,498	0,099	0,097	0,493	0,493	0,231	0,202	0,756	0,497
22	0,914	0,867	0,413	0,618	0,127	0,102	0,606	0,601	0,171	0,152	0,817	0,696
23	0,952	0,930	0,310	0,489	0,181	0,173	0,545	0,521	0,315	0,267	0,672	0,511
24	0,973	0,960	0,295	0,231	0,113	0,100	0,626	0,654	0,379	0,398	0,548	0,409
25	0,989	0,976	0,176	0,114	0,130	0,097	0,675	0,651	0,436	0,443	0,534	0,399
26	0,994	0,991	0,249	0,210	0,184	0,194	0,744	0,797	0,511	0,520	0,435	0,300
27	0,997	0,996	0,387	0,329	0,142	0,150	0,819	0,850	0,374	0,343	0,466	0,398
28	0,999	0,998	0,407	0,390	0,158	0,179	0,856	0,887	0,364	0,315	0,313	0,217
29	1,000	1,000	0,615	0,581	0,248	0,278	0,801	0,788	0,392	0,350	0,159	0,224
30	1,000	1,000	0,385	0,360	0,235	0,293	0,868	0,856	0,253	0,169	0,227	0,302
31	1,000	1,000	0,398	0,420	0,246	0,328	0,775	0,761	0,313	0,269	0,250	0,155
32	1,000	1,000	0,598	0,601	0,351	0,484	0,537	0,525	0,395	0,344	0,127	0,066
33	1,000	1,000	0,655	0,732	0,622	0,713	0,367	0,322	0,174	0,147	0,214	0,139
34	1,000	1,000	0,773	0,772	0,878	0,915	0,341	0,350	0,244	0,267	0,258	0,180
35	1,000	1,000	0,716	0,782	0,882	0,919	0,349	0,300	0,240	0,261	0,248	0,192
36	1,000	1,000	0,888	0,905	0,676	0,662	0,478	0,398	0,271	0,324	0,175	0,150
37	1,000	1,000	0,980	0,990	0,858	0,843	0,729	0,778	0,036	0,030	0,037	0,023
38	1,000	1,000	1,000	1,000	0,781	0,671	0,715	0,764	0,109	0,105	0,000	0,000

Legenda: 1 = PROFMAT - UFSJ e 2 = DEMAT - UFMG.

Fonte: Próprio autor e <http://www.mat.ufmg.br/futebol>.

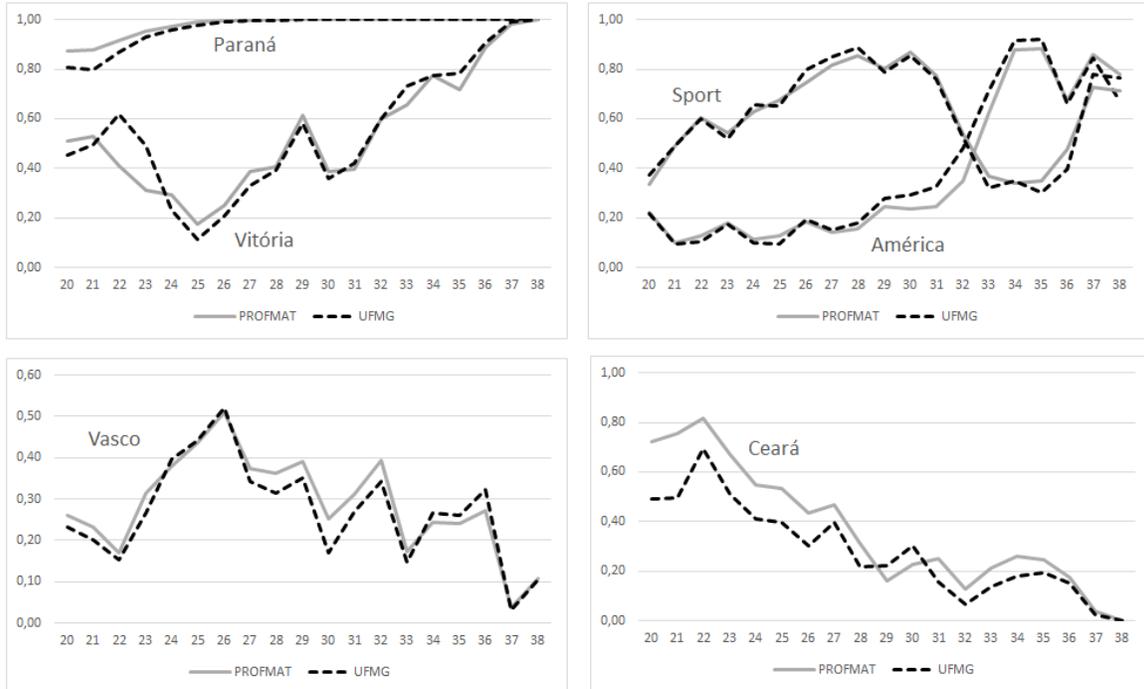


Figura 6: Probabilidade de rebaixamento para os seis piores times do campeonato
 Fonte: Próprio autor e <http://www.mat.ufmg.br/futebol/>

4 Considerações Finais

Neste artigo realizamos, através da distribuição de Poisson, uma aplicação de uma metodologia bem simples para calcular probabilidades no futebol. Aplicando a 250 jogos da Série A do Campeonato Brasileiro de 2018, a metodologia estudada apresentou bons resultados preditivos, a média da medida de DeFinetti foi 0,568 e aproximadamente 57% dos resultados estão abaixo de $2/3$.

Com a simulação foi possível prever com boa precisão o campeão (Palmeiras), e obteve resultados satisfatórios para os classificados para a Copa Libertadores da América, como também para os times rebaixados. Apesar de 250 jogos ser uma amostra pequena e da falta de acerto de empate e de vitória do visitante, os resultados obtidos pelo o método do PROFMAT estão coerentes com o modelo do Departamento de Matemática da UFMG, que é um modelo de referência, e respeitado nacionalmente, além disso o modelo PROFMAT (com distribuição de Poisson para os gols) é mais indicado para modelar torneios em mata-mata, enquanto o da UFMG mostra-se mais “estável” para pontos corridos. Isso deve-se ao fato que o nosso modelo PROFMAT baseia a

predição do resultado de cada partida no placar, enquanto o modelo UFMG sorteia qual o resultado da partida – vitória, empate ou derrota para o mandante. Entretanto, excetuando-se raríssimos casos em que seja necessário o desempate pelo saldo de gols¹, o placar da partida em si pouco importa. Já em um torneio em mata-mata, como a Copa do Brasil, o placar é fundamental para decidir quem se classifica para a próxima fase.

Para trabalhos futuros é necessário melhorar a qualidade do ajuste para prever empates e vitórias do visitante, além disso, uma linha de pesquisa futura é aplicar a média de gols para os últimos 5 ou 10 jogos e não para o campeonato todo. Também seria interessante colocar um peso para os gols, por exemplo, fazer gol no Palmeiras, que teve a melhor defesa em 2018, deveria ter um peso maior do que fazer gol no Paraná, que foi a pior defesa. Em caso de clássicos, os dois times podem entrar como mandante, já que geralmente clássicos não tem favoritos.

A metodologia apresentada pode ser usada para outros campeonatos de pontos corridos, como por exemplo a Série B do Campeonato Brasileiro ou o Campeonato Inglês. É importante ressaltar que o modelo utilizado neste trabalho não leva em consideração vários fatores que podem influenciar o resultado de um determinado jogo, tais como a presença de torcida, suspensões, entre outros. Todas as implementações computacionais foram realizadas utilizando o Software R que é um ambiente computacional de livre acesso. Os códigos podem ser encontrados através do link <https://github.com/luizpramos/prevCBA>.

Referências

- [1] B. R. James, *Probabilidade: um curso em nível intermediário*. IMPA, 2015.
- [2] B. N. B. de Lima, F. Brochero, G. N. Costa, G. M. Zeferino, M. T. Cunha e R. V. Martins: *Probabilidades no futebol*. Revista Matemática Universitária, ns.48/49, artigo 02, junho/dezembro de 2010.
- [3] A. K. Suzuki: *Modelagem Estatística para Previsão Esportiva: Uma aplicação no Futebol*. Dissertação de Mestrado, UFSCAR, 2007.
- [4] F. F. Farias: *Análise e Previsão de Resultados de Partidas de Futebol*. Dissertação de Mestrado, UFRJ, 2008.

¹Desde 2003 o campeonato brasileiro da série A é disputado na forma de pontos corridos, e nunca houve um empate em pontos na disputa do título, por exemplo.

- [5] L. Arruda: *Poisson, Bayes, Futebol e DeFinetti*. Dissertação de Mestrado, USP, 2000.
- [6] A. Suzuki e L. Tavares: *Modelagem Estatística para Previsão Esportiva: Uma aplicação no Futebol*. Revista Eletrônica Matemática e Estatística em Foco, Volume 3 - Número 1, Páginas: 32 a 47, Maio de 2015.
- [7] B. DeFinetti: *Probability, Induction and Statistics*. John Wiley, London, 1972.
- [8] J.B. Keller: *A Characterization of the Poisson Distribution and the Probability of Winning a Game*. The American Statistician, 48(4):294–298, 1994.
- [9] A. Lee: *Modeling Scores in the Premier League: Is Manchester United Really the Best?* Chance, 10(1):15–19, 1997.
- [10] R. Pollard: *Home advantage in soccer: a retrospective analysis*. Journal of Sports Sciences, pp. 237 – 248, 1986.
- [11] A. Suzuki, L. Salazar, F. Louzada-Neto e J. Leite: *A bayesian approach for predicting match outcomes: The 2006 (Association) Football World Cup*. Journal of the Operational Research Society, 61:1530–1539 (October 2010), 2009.

Submetido em 11 de Maio de 2020.

Aceito em 04 de Fevereiro de 2021.