UNIVERSIDADE ESTADUAL DA PARAÍBA
CAMPUS I
CENTRO DE CIÊNCIAS E TECNOLOGIA
DEPARTAMENTO DE ESTATÍSTICA
CURSO DE BACHARELADO EM ESTATÍSTICA

ROGÉRIO SILVA MORATO HERCULANO

CÁLCULOS DE PROBABILIDADE NOS JOGOS SEMIFINAIS E FINAIS DO CAMPEONATO PARAIBANO DE FUTEBOL 2020

CAMPINA GRANDE - PB
2021

ROGÉRIO SILVA MORATO HERCULANO

CÁLCULOS DE PROBABILIDADE NOS JOGOS SEMIFINAIS E FINAIS DO CAMPEONATO PARAIBANO DE FUTEBOL 2020

Trabalho de Conclusão de Curso (Artigo) apresentado ao curso de Bacharelado em Estatística do Departamento de Estatística do Centro de Ciências e Tecnologia da Universidade Estadual da Paraíba, como requisito parcial à obtenção do título de bacharel em Estatística.

Orientador: Prof. Gustavo Henrique Esteves

CAMPINA GRANDE - PB
2021

H539c Herculano, Rogério Silva Morato.

Cálculos de probabilidade nos jogos semifinais e finais do Campeonato Paraibano de Futebol 2020 [manuscrito] / Rogerio Silva Morato Herculano. - 2021.

20 p. : il. colorido.

Digitado.

Trabalho de Conclusão de Curso (Graduação em Estatística) - Universidade Estadual da Paraíba, Centro de Ciências e Tecnologia , 2021.

"Orientação : Prof. Dr. Gustavo Henrique Estreves , Departamento de Estatística - CCT."

1. Distribuição de Holgate. 2. Probabilidade. 3. Futebol. I. Título

21. ed. CDD 519.2

Elaborada por Giulianne M. Pereira - CRB - 15/714 BC/UEPB

ROGÉRIO SILVA MORATO HERCULANO

CÁLCULOS DE PROBABILIDADE NOS JOGOS SEMIFINAIS E FINAIS DO CAMPEONATO PARAIBANO DE FUTEBOL 2020

Trabalho de Conclusão de Curso (Artigo) apresentado ao curso de Bacharelado em Estatística do Departamento de Estatística do Centro de Ciências e Tecnologia da Universidade Estadual da Paraíba, como requisito parcial à obtenção do título de bacharel em Estatística.

Trabalho aprovado em 19 de fevereiro de 2021.

BANCA EXAMINADORA

Prof. Gustavo Henrique Esteves (Orientador)
Universidade Estadual da Paraíba (UEPB)

Prof. Dr. Tiago Almeida de Oliveira
Universidade Estadual da Paraíba (UEPB)

Prof. Dr. Kleber Napoleão Nunes de Oliveira Barros
Universidade Federal Rural de Pernambuco (UFRPE)

Dedico esse trabalho em especial a minha mãe, a toda a minha família e aqueles que me ajudaram chegar até aqui, só tenho palavras de gratidão, que Deus continue os abençoando e protegendo sempre, meu muito obrigado por tudo.

LISTA DE ILUSTRAÇÕES

Figura 1 – Gráfico de barras com as probabilidades (em porcentagem) de classificação às finais. 13
Figura 2 – Gráfico de barras com as probabilidades (em porcentagem) de conquista do título. 15

LISTA DE TABELAS

Tabela 1 – Probabilidades, em porcentagem, calculadas para os jogos da primeira semifinal. 12
Tabela 2 – Probabilidades, em porcentagem, calculadas para os jogos da segunda semifinal. 12
Tabela 3 – Probabilidades de classificação à final, de acordo com cada semifinal. . 13
Tabela 4 – Probabilidades, em porcentagem, calculadas para os jogos da final do Campeonato Paraibano de 2020. 14

SUMÁRIO

1 INTRODUÇÃO 8

2 METODOLOGIA 9
2.1 Base de dados 9
2.2 Métodos estatísticos 9
2.2.1 Distribuição bivariada de Holgate *9*
2.2.2 Estimação dos parâmetros *10*
2.2.3 Simulação *11*

3 RESULTADOS E DISCUSSÃO 11
3.1 Jogos das semifinais 12
3.2 Jogos das finais 14

4 CONCLUSÃO 15

REFERÊNCIAS 16

APÊNDICE A – CÓDIGOS USADOS NO TRABALHO 17
A.1 Códigos usados nos cálculos das semifinais 17
A.2 Códigos usados nos cálculos das finais 18

CÁLCULOS DE PROBABILIDADE NOS JOGOS SEMIFINAIS E FINAIS DO CAMPEONATO PARAIBANO DE FUTEBOL 2020

PROBABILITY CALCULATIONS IN THE SEMIFINAL AND FINAL MATCHES OF 2020 PARAIBANO SOCCER CHAMPIONSHIP

Rogério Silva Morato Herculano*
Gustavo H. Esteves†

RESUMO

Este trabalho teve por finalidade estimar as probabilidades de que os clubes classificados para as semifinais do Campeonato Paraibano de Futebol de 2020 conseguissem passar à final e de conquistarem o título do campeonato. Foram determinadas as probabilidades de cada clube semifinalista conquistar a vaga na final e suas respectivas probabilidades de conquista do título através de um método estatístico conhecido como Soma e Diferença, baseado na distribuição bivariada de Poisson de Holgate, além de simulação computacional dos jogos de futebol. O Botafogo apresentou a maior probabilidade de conquista do campeonato enquanto o Treze teve a menor. Entretanto, a final foi realizada entre Campinense e Treze, sendo estimada uma probabilidade de conquista pelo Campinense maior do que a do rival. Mesmo com uma probabilidade inferior de conquista em relação aos demais clubes postulantes ao título, o Treze obteve êxito na final da competição, demonstrando que nem sempre uma probabilidade maior implica a ocorrência do evento. Um clube com a probabilidade maior pode não conquistar o título, o que foi evidenciado neste trabalho.

Palavras-chaves: Distribuição de Holgate. Soma e Diferença. Futebol.

ABSTRACT

This work had the purpose of estimating the probabilities of the teams that qualified for the semi-finals of the Paraibano Soccer Championship 2020 to be able to pass to the final and to become the winner of the championship. The probabilities of each semifinalist club winning the spot in the final and their respective likelihoods of winning the tournament were determined using a statistical method known as Sum and Difference based on the bivariate distribution of Holgate's Poisson, in addition to computer simulation of soccer matches. Botafogo had the highest probability of winning the championship while Treze had the lowest one. However, the final matches was held between Campinense and Treze, with a higher probability of conquest by Campinense than that for the rival. Even with a lower probability of conquest in relation to the other clubs applying for the title, Treze was successful in the final of the competition, demonstrating that not always a higher probability implies the occurrence of the event. A club with the highest probability may not win the title, as evidenced in this work.

Keywords: Holgate's Distribution. Sum and Diference. Soccer.

* Aluno do curso de Estatística, Departamento de Estatística, UEPB, Campina Grande, PB.
† Professor Doutor Associado A, Departamento de Estatística, UEPB, Campina Grande, PB, gesteves@servidor.uepb.edu.br

1 INTRODUÇÃO

O futebol é um esporte popular em todo o planeta, não sendo nenhum abuso dizer que ele é o mais popular dentre todos os esportes. Isso pode ser constatado pela repercussão da Copa do Mundo de Futebol, que é realizada a cada quatro anos. Esta realidade não é diferente no Brasil, onde os campeonatos nacionais e estaduais sempre movimentam torcedores e imprensa de todas as regiões do país.

O Campeonato Paraibano de Futebol é a competição organizada pela Federação Paraibana de Futebol (FPF) para a disputa do título estadual entre os clubes da Paraíba. Ele foi disputado pela primeira vez em 1908 e é um dos quatro campeonatos estaduais mais antigos do Brasil, no entanto até 1938 não contava com equipes do interior, assim as vagas eram apenas para as equipes da capital do estado, João Pessoa-PB. Atualmente conta com dez equipes na primeira divisão e promove os clubes mais bem classificados para a Copa do Nordeste, Copa do Brasil e a Série D do Campeonato Brasileiro.

No ano de 2020 aconteceu a 110ª edição do Campeonato Paraibano de Futebol (a maior liga profissional do esporte no estado), que teve início no dia 21 de janeiro de 2020 e terminou no dia 15 de agosto do mesmo ano, mas cabe salientar que a temporada foi suspensa em 18 de março, devido a pandemia da Covid-19, sendo retomada em 18 de julho. O campeão foi o Treze Futebol Clube, sendo rebaixados os clubes CSP e Sport-PB, ao todo foram disputadas 56 partidas ao longo da competição com 128 gols marcados, o que dá uma média de 2,29 por jogo, o artilheiro da competição foi Rafael Ibiapino do Campinense Clube, com 9 gols marcados.

Neste contexto, este trabalho teve por objetivo verificar e mensurar as possibilidades que cada um dos quatro clubes classificados na fase de grupos do Campeonato Paraibano de Futebol 2020 tinham de passar para a final e posteriormente, uma vez definidos os finalistas, calcular as probabilidades de cada um dos dois se tornar campeão estadual no ano. Para estes cálculos aqui realizados, foi utilizada uma família de distribuições bivariadas de Poisson para modelagem de jogos de futebol, além de um método de estimação de parâmetros conhecido como Soma e Diferença (SD).

Este método conta com uma alternativa mais simples que não estima a covariância da distribuição bivariada de Poisson, admitindo-a nula, sendo conhecido como SD0. Porém, neste trabalho optou-se por utilizar uma versão do método conhecida como SD1, que inclui a estimação da covariância do modelo probabilístico bivariado (λ_{12}).

Divididos em dois grupos na competição, os clubes que se classificaram na primeira fase foram Treze e Botafogo pelo grupo A, e Campinense e Sousa pelo grupo B. De acordo com as regras do torneio, as semifinais foram disputadas em jogos de ida e volta entre os dois clubes classificados de cada grupo. Devido ao histórico de resultados recentes e maior poder de investimento, seria razoável pensar que tanto Botafogo como Campinense seriam os clubes mais propensos a chegarem à grande final do campeonato, algo que a teoria de probabilidades corroborou, como será apresentado mais a frente.

Porém, após jogos muito equilibrados nas semifinais houve a necessidade das cobranças de pênaltis para definir quais clubes se classificariam, a final ficou definida entre Treze e Campinense, dois grandes rivais locais da cidade de Campina Grande-PB. O Campinense também apresentava um certo favoritismo probabilístico, mas após a longa paralisação do campeonato em função da pandemia da Covid-19, no mês de agosto o Treze Futebol Clube sagrou-se campeão Paraibano de 2020 ao derrotar o Campinense Clube na final.

2 METODOLOGIA

Neste trabalho foram utilizadas técnicas estatísticas baseadas em uma família de distribuições bivariadas de Poisson e um particular método de estimação dos parâmetros associados ao modelo estocástico. Tais estimativas possibilitam cálculos de probabilidades associadas aos possíveis resultados de jogos de futebol, que neste contexto foram aplicados aos jogos semifinais e finais do Campeonato Paraibano de 2020. A seguir estão descritos os dados utilizados neste trabalho e um breve resumo dos métodos adotados.

2.1 Base de dados

Os dados utilizados correspondem aos resultados de todos os jogos de futebol, na forma do placar final com os números de gols marcados por ambas equipes, realizados no Campeonato Paraibano de 2019 juntamente com todos os jogos da primeira fase do Campeonato Paraibano de 2020, para os cálculos de probabilidades envolvendo os jogos semifinais. Para os cálculos envolvendo as finais do campeonato, também foram incluídos os resultados destes jogos semifinais de 2020.

O Campeonato Paraibano de Futebol de 2019 foi composto por 10 clubes e teve início em janeiro do mesmo ano com um sistema de turno e returno na primeira fase, que classificou quatro times para as semifinais. Os dois vencedores das semifinais fizeram a final. Ao todo aconteceram 56 jogos em toda a competição, sendo 50 jogos na primeira fase, quatro jogos semifinais e dois jogos finais. O mesmo sistema de disputa foi adotado no ano de 2020, de modo que para os cálculos de probabilidades dos jogos semifinais deste ano foram utilizados 106 jogos para a estimação dos parâmetros. Já para os cálculos envolvendo as finais, foram utilizados 110 jogos, com a inclusão dos resultados das semifinais de 2020.

2.2 Métodos estatísticos

A distribuição de probabilidades adotada foi a distribuição bivariada de Poisson de Holgate (HOLGATE, 1964). Segundo Arruda (2000) esta é a melhor opção dentre as famílias de distribuições bivariadas de Poisson para modelar jogos de futebol, fato corroborado também por Suzuki (2007). Maiores informações sobre probabilidade, entre elas, as famílias de distribuições bivariadas de Poisson podem ser encontradas em Johnson, Kemp e Kotz (2005).

Toda abordagem teórica utilizada neste trabalho foi estudada e relatada por Silva (2019), que apresenta uma revisão teórica detalhada dos métodos envolvidos e também fez um trabalho prático envolvendo estimativas de parâmetros de todos os jogos do Campeonato Brasileiro de 2019, com resultados interessantes.

2.2.1 Distribuição bivariada de Holgate

A família de distribuições bivariadas de Holgate pode ser construída a partir de três processos de Poisson independentes, sendo P_1,P_{12} e P_2 respectivamente com taxas λ_1, λ_{12} e λ_2, de modo que se possa escrever duas variáveis aleatórias tais que $X_1 = P_1 + P_{12}$ e $X_2 = P_2 + P_{12}$. Assim, cada uma das variáveis X_1 e X_2 têm distribuição univariada de Poisson e conjuntamente seguem a distribuição bivariada de Holgate, com covariância dada por λ_{12}, que é a taxa do processo P_{12} (ARRUDA, 2000; SILVA, 2019; SUZUKI, 2007). Assim, de acordo com Arruda (2000) a função de probabilidade de (X_1, X_2) pode

ser escrita da seguinte forma:

$$P(X_1 = x_1, X_2 = x_2) = e^{-(\lambda_1+\lambda_2+\lambda_{12})} \sum_{i=0}^{\min(x_1,x_2)} \frac{\lambda_1^{x_1-i}\lambda_2^{x_2-i}\lambda_{12}^{i}}{(x_1-i)!(x_2-i)!i!}.$$

No contexto da modelagem dos jogos de futebol, os processos de Poisson P_1 e P_2 estão relacionados com a capacidade de cada um dos times mandante e visitante, respectivamente, marcarem seus gols. Por outro lado, o processo P_{12} está relacionado com fatores do jogo que podem influenciar simultaneamente nesta capacidade dos times marcarem seus gols, tais como fator campo, condições climáticas, etc. E finalmente as variáveis X_1 e X_2 modelam os números de gols marcados pelos times mandante e visitante, respectivamente.

2.2.2 Estimação dos parâmetros

Para a estimação dos parâmetros envolvidos com a distribuição bivariada de Holgate, foi utilizado um método conhecido como Soma e Diferença (SD). Neste método existem duas versões, uma mais simples (SD0) que assume independência entre os números de gols marcados pelos dois times que disputam uma partida de futebol, e outra que não faz tal suposição (SD1).

De acordo com Silva (2019) os métodos de estimação SD se baseiam na seguintes propriedades:

$$\begin{aligned} E(X_1 - X_2) &= \lambda_1 - \lambda_2 \\ E(X_1 + X_2) &= \lambda_1 + \lambda_2 + 2\lambda_{12}. \end{aligned}$$

Supor independência entre X_1 e X_2 equivale a $\lambda_{12} = 0$, o que simplifica o processo de estimação. Porém, no contexto dos jogos de futebol esta suposição não é realista, de modo que aqui optou-se por usar o método SD1.

Para a estimação do parâmetro da covariância no método SD1, além das duas propriedades acima, uma terceira se faz necessária que é dada pela seguinte expressão

$$E[(X_1 + X_2)^2] - [E(X_1 + X_2)]^2 = \lambda_1 + 4\lambda_{12} + \lambda_2.$$

A partir do sistema constituído por estas três propriedades que envolvem os valores esperados para a soma $(X_1 + X_2)$ e a diferença $(X_1 - X_2)$ dos gols marcados pelos times mandante e visitante é possível se escrever expressões para os três parâmetros envolvidos. Assim, substituindo estas esperanças por suas respectivas estimativas, chega-se aos estimadores dos parâmetros:

$$\begin{cases} \hat{\lambda}_1 = \frac{\hat{E}(X_1-X_2)+2\hat{E}(X_1+X_2)-\left\{\hat{E}[(X_1+X_2)^2]-[\hat{E}(X_1+X_2)]^2\right\}}{2} \\ \hat{\lambda}_2 = \frac{2\hat{E}(X_1+X_2)-\hat{E}(X_1-X_2)-\left\{\hat{E}[(X_1+X_2)^2]-[\hat{E}(X_1+X_2)]^2\right\}}{2} \\ \hat{\lambda}_{12} = \frac{\left\{\hat{E}[(X_1+X_2)^2]-[\hat{E}(X_1+X_2)]^2\right\}-\hat{E}(X_1+X_2)}{2} \end{cases}.$$

Os valores esperados envolvidos com os cálculos, ou seja, $E[X_1 + X_2]$, $E[X_1 - X_2]$ e $E[(X_1 + X_2)^2]$, são estimados através de modelos lineares definidos por

$$\begin{aligned} (X_1 + X_2)_i &= \mathbf{G_i}\alpha + \epsilon_{ai} \\ (X_1 - X_2)_i &= \mathbf{H_i}\beta + \epsilon_{bi} \\ \left[(X_1 + X_2)^2\right]_i &= \mathbf{G_i}\gamma + \epsilon_{ci}, \end{aligned}$$

onde $i = 1, 2, ..., n$; com n denotando o número de jogos da base de dados; ϵ_{ai}, ϵ_{bi} e ϵ_{ci} são erros independentes com médias iguais a 0. **G** e **H** são matrizes onde as linhas representam os jogos e as colunas os times de futebol, em ambos os casos a última coluna é preenchida por zeros (0) ou uns (1), onde o valor um representa se houve fator de campo no jogo. Na matriz **G** cada linha é preenchida por zeros, para os times que não participaram do jogo, e uns para os dois times que jogaram aquela partida. A mesma ideia se aplica à matriz **H**, porém o time mandante assume o valor 1 e o visitante o valor -1. Os termos α, β e γ denotam vetores de parâmetros para todos os times.

Existem ainda outras famílias de métodos para estimação de parâmetros para este tipo de modelagem. Dentre eles, por exemplo, existem os métodos baseados em modelos *log*-lineares, também conhecidos como métodos de chances (LEE, 1997), ou mesmo métodos baseados em abordagens bayesianas (SUZUKI et al., 2017). Porém, como não foram usados neste trabalho, não serão abordados aqui.

2.2.3 Simulação

A partir dos dados referentes aos resultados dos jogos do Campeonato Paraibano dos anos de 2019 e 2020 (primeira fase) foi usado o método SD1 apresentado na seção anterior para se estimar os parâmetros de cada um dos jogos das semifinais do campeonato de 2020, que foram Campinense × Sousa e Treze × Botafogo, ambos em confrontos de ida e volta. Uma vez calculadas estas estimativas, as probabilidades de vitória do mandante, empate ou vitória do visitante foram calculadas a partir da distribuição bivariada de Poisson de Holgate. A mesma ideia foi usada para os dois jogos da grande final, com a inclusão dos resultados das semifinais na base de dados.

Dentre os quatro clubes semifinalistas foram calculadas as probabilidades frequentistas de classificação de cada um deles para a final, com base na simulação computacional de todos os quatro jogos. Neste sentido, um milhão de resultados destes jogos foram gerados computacionalmente com base na distribuição de Poisson de Holgate com os parâmetros estimados, conforme especificado acima. O mesmo procedimento foi adotado para calcular as probabilidades de conquista do título entre os dois finalistas.

Todos os cálculos foram executados no programa de computação estatística R (R Core Team, 2020). Para os cálculos de probabilidades envolvendo a modelagem probabilística bivariada de Holgate foi utilizado o pacote *extraDistr* (WOLODZKO, 2020) e para a estimação dos parâmetros do modelo foi utilizado o pacote experimental *socceR*[1] (ESTEVES, 2020). A seguir serão apresentados os resultados obtidos.

3 RESULTADOS E DISCUSSÃO

A partir dos métodos apresentados na seção anterior foi possível calcular as probabilidades envolvidas com os jogos das semifinais do Campeonato Paraibano de 2020. Também foram calculadas através de simulação computacional as probabilidades de cada uma das equipes conseguirem se classificar para a final da competição. As duas semifinais foram Campinense × Sousa e Treze × Botafogo.

Após os jogos realizados, passaram à final Treze e Campinense. De modo semelhante ao que foi feito nos jogos das semifinais, as probabilidades envolvidas com os dois jogos da final também foram calculadas, e as probabilidades de conquista do título foram calculadas por simulação.

[1] <https://github.com/ghesteves/socceR>

3.1 Jogos das semifinais

A primeira semifinal ficou definida entre os times do Campinense e do Sousa, com o Campinense decidindo o segundo jogo com o seu mando de campo. A Tabela 1 apresenta as probabilidades calculadas para os dois jogos desta semifinal.

Tabela 1 – Probabilidades, em porcentagem, calculadas para os jogos da primeira semifinal.

Jogo	Vitória Mandante	Empate	Vitória Visitante
Sousa × Campinense	36,3	20,9	42,8
Campinense × Sousa	59,2	18,8	22,0

Fonte: Produzida pelos autores.

Na Tabela 1 observa-se que para o primeiro jogo, realizado na cidade de Sousa-PB, o Campinense apresentou uma probabilidade maior de vitória com 42,8%, mesmo jogando fora de casa. No entanto, o resultado final da partida foi empate em 2 a 2, que era o resultado menos provável nas estimativas. Na segunda partida realizada na cidade de Campina Grande-PB, o resultado final da partida foi um novo empate, mas desta vez por 0 a 0, onde houve a necessidade de realização de cobranças de pênaltis para definir o Campinense como primeiro finalista do Campeonato Paraibano de Futebol. Segundo as estimativas, os resultados menos prováveis aconteceram nos dois jogos, o que mostra que nem sempre uma maior probabilidade implica na ocorrência do evento.

A segunda semifinal ficou definida entre Treze e Botafogo, com o Treze decidindo o segundo confronto em casa. A Tabela 2 apresenta as probabilidades calculadas para os dois jogos desta semifinal.

Tabela 2 – Probabilidades, em porcentagem, calculadas para os jogos da segunda semifinal.

Jogo	Vitória Mandante	Empate	Vitória Visitante
Botafogo × Treze	68,0	22,6	9,4
Treze × Botafogo	25,1	28,8	46,1

Fonte: Produzida pelos autores.

Nesta Tabela 2 observa-se que para o primeiro confronto entre Botafogo e Treze, realizado na cidade de João Pessoa-PB, as probabilidades estimadas apontaram uma ampla vantagem para o clube da capital com 68% de vitória, contra apenas 9,4% para o Treze. Este favoritismo probabilístico de fato se confirmou e o Botafogo venceu a partida por 2 a 0 no estádio Almeidão. Já no segundo confronto realizado na cidade de Campina Grande-PB, o Treze ainda apresentava uma probabilidade menor de vitória, com 25,1%, sendo inferior inclusive à de empate (28,8%), porém, o resultado final da partida foi 2 a 0 para o Treze, que devolveu a derrota sofrida no primeiro confronto e acabou se classificando para a final após a disputa nos pênaltis. Estes resultados mostram que as probabilidades estimadas apresentavam uma ampla vantagem para o Botafogo no primeiro jogo e ainda uma ligeira vantagem no segundo jogo. A vitória do clube da capital confirmou este favoritismo no confronto inicial, porém o Treze conseguiu usar seu fator campo para reverter o confronto

e garantir participação na final do campeonato em um resultado que pode ser considerado improvável.

Neste contexto, as probabilidades de classificação à final dos quatro clubes que disputaram as semifinais foram calculadas computacionalmente. Para isso, os resultados dos quatro jogos das semifinais foram simulados um milhão de vezes no programa R e as probabilidades foram estimadas através das frequências relativas de vitória de cada clube nos dois jogos da sua respectiva semifinal.

Tabela 3 – Probabilidades de classificação à final, de acordo com cada semifinal.

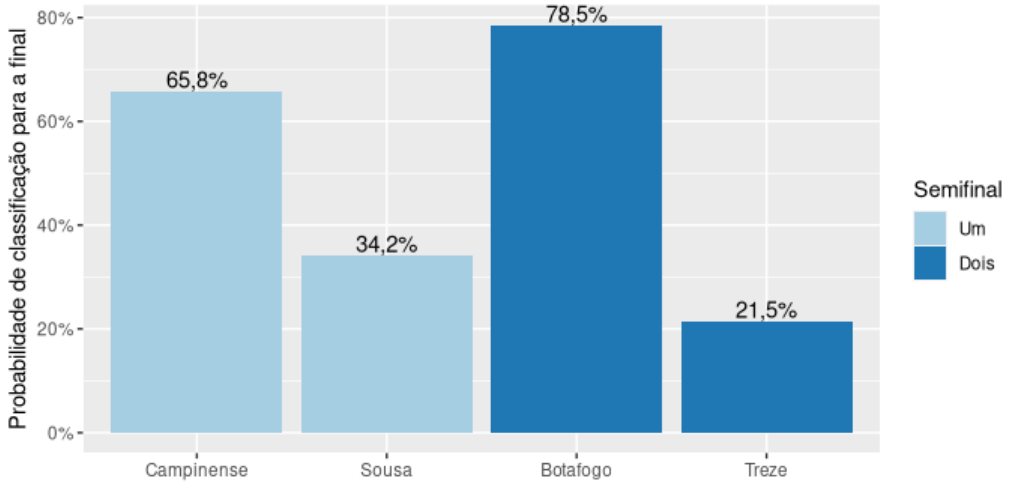
Disputa	**Time**	**Probabilidade (em %)**
Semifinal 1	Campinense	65,8
	Sousa	34,2
Semifinal 2	Botafogo	78,5
	Treze	21,5

Fonte: Produzida pelos autores.

A Tabela 3 apresenta estas probabilidades de classificação à final calculadas por simulação. É possível se observar que o Botafogo apresentou uma probabilidade de 78,5% de classificação contra 21,5% do Treze, que foi o seu adversário na semifinal. Por outro lado, o Campinense apresentou 65,8% de probabilidade de classificação contra 34,2% do Sousa, seu adversário na disputa. Em outras palavras, estas probabilidades estimadas indicam que pode-se afirmar que Campinense e Botafogo eram probabilisticamente favoritos em seus respectivos confrontos.

De um lado o Campinense confirmou este favoritismo, mas do outro o Treze conseguiu reverter uma probabilidade amplamente desfavorável e se classificar para disputar a final contra seu rival local. Tais resultados mostram que embora a final mais provável seria disputada entre Botafogo e Campinense, isso não aconteceu. Estes valores de probabilidades também estão representados graficamente na Figura 1.

Figura 1 – Gráfico de barras com as probabilidades (em porcentagem) de classificação às finais.



Vale ressaltar que dentre as quatro equipes semifinalistas o Treze era quem tinha a menor probabilidade de se classificar para a final (Tabela 3 e Figura 1), enquanto o Botafogo chegou às semifinais com a maior probabilidade de conquista, mas acabou eliminado nos pênaltis. No outro confronto o Sousa chegou um pouco desacreditado, com um elenco reduzido e baixa folha salarial, o que refletiu em uma maior probabilidade de classificação do Campinense, o que de fato aconteceu, muito embora sem facilidade dado que a decisão também se deu nos pênaltis.

3.2 Jogos das finais

Uma vez definido o confronto final entre Treze e Campinense, também foram calculadas as probabilidades para os dois jogos desta disputa, que aconteceu no estádio Amigão, na cidade de Campina Grande-PB. Neste caso, os resultados dos jogos das semifinais foram inseridos nos dados usados para os cálculos. Tais valores de probabilidades estão apresentados na Tabela 4.

Tabela 4 – Probabilidades, em porcentagem, calculadas para os jogos da final do Campeonato Paraibano de 2020.

Jogo	**Vitória Mandante**	**Empate**	**Vitória Visitante**
Campinense × Treze	56,2	23,6	20,2
Treze × Campinense	38,8	26,0	35,2

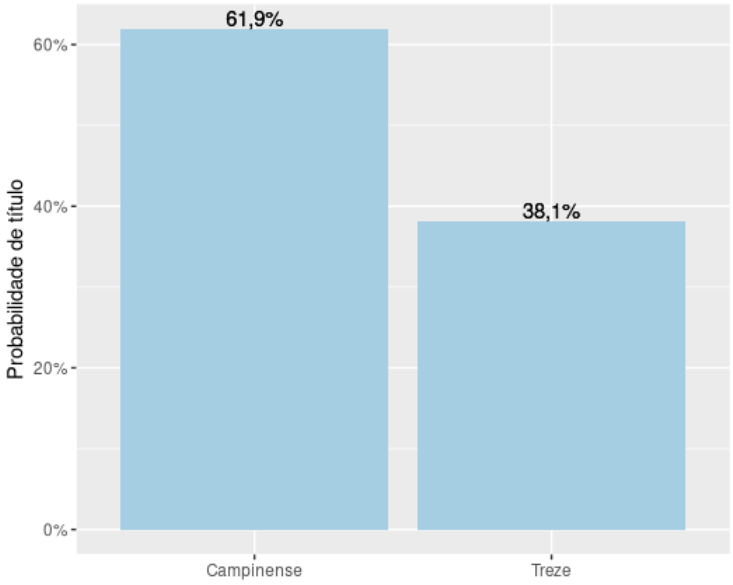
Fonte: Produzida pelos autores.

De acordo com a Tabela 4 o Campinense apresentou maior probabilidade de vitória na primeira partida da final, 56,2% contra 20,2% do seu rival, mas o resultado foi vitória do Treze por 2 a 0. No segundo confronto o Treze apresentava uma probabilidade de vitória ligeiramente maior, 38,8% contra 35,2% do rival, mas desta vez o Campinense venceu a partida por 1 a 0. Da mesma forma que nas semifinais, os dois jogos da final foram simulados um milhão de vezes no programa R para se calcular as probabilidades de conquista do título entre os dois rivais. Novamente o Treze estava em desvantagem nestas estimativas, com 38,1% contra 61,9% do Campinense, dados representados na Figura 2. Porém, na soma dos resultados dos dois jogos o Treze sagrou-se campeão estadual de 2020 mesmo com esta estimativa de probabilidade menor de consegui-lo.

Os métodos utilizados aqui foram capazes de calcular probabilidades tanto para os jogos das semifinais como da final do Campeonato Paraibano de Futebol de 2020, onde ficou claro que não necessariamente um clube com maior probabilidade de vitória em um jogo pode de fato ganhá-lo. Outros fatores podem colaborar para que o resultado mais provável não aconteça, da mesma forma que o Treze se sagrou campeão na final, mesmo sendo o desfecho menos provável.

Também é interessante citar que Ntzoufras (2009) apresenta um exemplo de cálculos probabilísticos usando esta mesma distribuição de Poisson bivariada para calcular probabilidades associadas a jogos de futebol da liga inglesa na temporada 2006-2007, porém, usando modelagem bayesiana.

Figura 2 – Gráfico de barras com as probabilidades (em porcentagem) de conquista do título.



4 CONCLUSÃO

Neste trabalho foram apresentados métodos para cálculos de probabilidades em jogos de futebol baseados na distribuição bivariada de Poisson de Holgate e em um método de estimação de parâmetros conhecido como Soma e Diferença (SD), mais especificamente o método SD1, que considera a estimação do parâmetro de covariância do modelo probabilístico. Em outro sentido foi usada simulação computacional para calcular estimativas frequentistas das probabilidades de cada equipe conseguir a classificação à final da competição bem como de conquista do título de cada equipe finalista.

Com relação aos resultados, as probabilidades dos clubes semifinalistas se classificarem para a final apontaram para a decisão entre Botafogo e Campinense. Porém o Treze conseguiu superar suas dificuldades, se classificando para disputar a final contra seu rival local. E na final, mesmo com os modelos indicando novamente uma menor probabilidade, o Treze Futebol Clube sagrou-se campeão estadual em 2020.

Estes resultados apontam para os cuidados que se deve tomar na interpretação de cálculos probabilísticos. Afinal, existiam fortes evidências baseadas nas estimativas calculadas de que a final do Campeonato Paraibano de 2020 seria disputada entre Botafogo e Campinense. Mesmo na final, após a surpresa da classificação do Treze, as probabilidades estimadas eram mais favoráveis para o Campinense. E mais uma vez o Treze conseguiu vencer.

Alguns tendem a argumentar que o modelo probabilístico errou, mas é importante se notar que em nenhum momento foi calculada uma estimativa de probabilidade zero para o Treze, tanto para a classificação à final como para a conquista do título. O que acontece é que existia uma probabilidade positiva do Treze se classificar e conquistar o título que, por mais improvável que fosse, poderia acontecer de fato.

REFERÊNCIAS

ARRUDA, M. L. d. *Poisson, Bayes, Futebol e DeFinetti*. Dissertação (Mestrado) — Universidade de São Paulo, 2000. Citado na página 9.

ESTEVES, G. H. *socceR: Probabilistic Models For Soccer Matches*. [S.l.], 2020. R package version 0.0.0.9000. Disponível em: <https://github.com/ghesteves/socceR>. Citado na página 11.

HOLGATE, P. Estimation for the bivariate poisson distribution. *Biometrika*, [Oxford University Press, Biometrika Trust], v. 51, n. 1/2, p. 241–245, 1964. ISSN 00063444. Disponível em: <http://www.jstor.org/stable/2334210>. Citado na página 9.

JOHNSON, N. L.; KEMP, A. W.; KOTZ, S. *Univariate discrete distributions*. [S.l.]: John Wiley & Sons, 2005. v. 444. Citado na página 9.

LEE, A. J. Modeling scores in the premier league: Is manchester united really the best? *Chance*, v. 10, n. 1, p. 15–19, 1997. Citado na página 11.

NTZOUFRAS, I. *Bayesian modeling using WinBUGS*. [S.l.]: John Wiley & Sons, 2009. Citado na página 14.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2020. Disponível em: <http://www.R-project.org/>. Citado na página 11.

SILVA, P. A. F. da. *Cálculo de probabilidades no futebol: uma aplicação no campeonato brasileiro de 2019*. 2019. Trabalho de Conclusão de Curso (TCC) — Universidade Estadual da Paraíba, 2019. Citado 2 vezes nas páginas 9 e 10.

SUZUKI, A. K. *Modelagem estatística para a determinação de resultados de dados esportivos*. Dissertação (Mestrado) — Universidade Federal de São Carlos, 2007. Citado na página 9.

SUZUKI, A. K. et al. Uma abordagem bayesiana para previsão de resultados de jogos de futebol: uma aplicação ao campeonato inglês. *Revista Brasileira de Biometria/Biometric Brazilian Journal*, v. 35, n. 1, p. 76–97, 2017. Citado na página 11.

WOLODZKO, T. *extraDistr: Additional Univariate and Multivariate Distributions*. [S.l.], 2020. R package version 1.9.1. Disponível em: <https://CRAN.R-project.org/package=extraDistr>. Citado na página 11.

APÊNDICE A – CÓDIGOS USADOS NO TRABALHO

Este apêndice apresenta os códigos do Programa R que foram usados nas análises apresentadas aqui, que está divido em duas seções, a primeira contendo o código utilizado para os cálculos das semifinais e segunda contendo o código utilizado para os cálculos das finais.

A.1 Códigos usados nos cálculos das semifinais

Esta seção apresenta o código fonte do R, usado para os cálculos relacionados com os jogos semifinais do Campeonato Paraibano de Futebol 2020.

```
## Carregando pacotes necessários
library(socceR)
library(extraDistr)

## Lendo o arquivo dos dados já carregados
load("dados_semis.RData")

## Calculando os coeficientes
coeficientes <- calcCoefSD1(dados)

## Número de réplicas das simulações
n <- 10^6


######### Primeira semifinal
### Sousa x Campinense (jogo 1)
tmp1 <- calcLambdas(coeficientes, "Sousa", "Campinense"); tmp1;
calcMatProb(tmp1)$Probs

## Fixando uma semente
set.seed(1)
simulJ1 <- rbvpois(n, tmp1[1], tmp1[2], tmp1[3])

#####  Campinense x Sousa (jogo 2)
tmp2 <- calcLambdas(coeficientes, "Campinense", "Sousa"); tmp2;
calcMatProb(tmp2)$Probs

## Fixando uma semente
set.seed(2)
simulJ2 <- rbvpois(n, tmp2[1], tmp2[2], tmp2[3])

## Calculando a soma dos placares
resFinal <- cbind(simulJ1[,1]+simulJ2[,2], simulJ1[,2]+simulJ2[,1])
## Calculando a diferença Sousa - Campinense
difFinal <- resFinal[,1] - resFinal[,2]

## Sorteando um vencedor ao acaso em caso de empate
idx <- which(difFinal == 0)
difFinal[idx] <- sample(c(-1,1), length(idx), rep=T)
```

```
probs <- NULL

## Probabilidade do Sousa ir pra final
probs[3] <- mean(difFinal > 0)
## Probabilidade do Campinense ir pra final
probs[2] <- mean(difFinal < 0)


######### Segunda semifinal
### Botafogo x Treze (jogo 1)
tmp1 <- calcLambdas(coeficientes, "Botafogo", "Treze"); tmp1;
calcMatProb(tmp1)$Probs

## Fixando uma semente
set.seed(3)
simulJ1 <- rbvpois(n, tmp1[1], tmp1[2], tmp1[3])

### Treze x Botafogo (jogo 2)
tmp2 <- calcLambdas(coeficientes, "Treze", "Botafogo"); tmp2;
calcMatProb(tmp2)$Probs

## Fixando uma semente
set.seed(4)
simulJ2 <- rbvpois(n, tmp2[1], tmp2[2], tmp2[3])

## Calculando a soma dos placares
resFinal <- cbind(simulJ1[,1]+simulJ2[,2], simulJ1[,2]+simulJ2[,1])
## Calculando a diferença Nacional - Botafogo
difFinal <- resFinal[,1] - resFinal[,2]

## Sorteando um vencedor ao acaso em caso de empate
idx <- which(difFinal == 0)
difFinal[idx] <- sample(c(-1,1), length(idx), rep=T)

## Probabilidade do Botafogo ir pra final
probs[1] <- mean(difFinal > 0)
## Probabilidade do Treze ir pra final
probs[4] <- mean(difFinal < 0)

names(probs) <- c("Botafogo", "Campinense", "Sousa", "Treze")
probs*100

barplot(probs*100, ylim=c(0,100))
```

A.2 Códigos usados nos cálculos das finais

Esta seção apresenta o código fonte do R, usado para os cálculos relacionados com os jogos finais do Campeonato Paraibano de Futebol 2020.

```
## Carregando pacotes necessários
library(socceR)
library(extraDistr)
```

```
## Lendo o arquivo dos dados já carregados
load("dados_finais.RData")

## Calculando os coeficientes
coeficientes <- calcCoefSD1(dados)

## Número de réplicas das simulações
n <- 10^6

########## Primeiro jogo da Final
###  Campinense x Treze
tmp1 <- calcLambdas(coeficientes, "Campinense", "Treze"); tmp1;
calcMatProb(tmp1)$Probs

## Fixando uma semente
set.seed(1)
simulJ1 <- rbvpois(n, tmp1[1], tmp1[2], tmp1[3])

########## Segundo jogo da Final
###  Treze x Campinense
tmp2 <- calcLambdas(coeficientes, "Treze", "Campinense"); tmp2;
calcMatProb(tmp2)$Probs

## Fixando uma semente
set.seed(2)
simulJ2 <- rbvpois(n, tmp2[1], tmp2[2], tmp2[3])

## Calculando a soma dos placares
resFinal <- cbind(simulJ1[,1]+simulJ2[,2], simulJ1[,2]+simulJ2[,1])
## Calculando a diferença Campinense - Atlético
difFinal <- resFinal[,1] - resFinal[,2]

## Sorteando um vencedor ao acaso em caso de empate
idx <- which(difFinal == 0)
difFinal[idx] <- sample(c(-1,1), length(idx), rep=T)

probs <- NULL

## Probabilidade do Campinense ser campeão
probs[1] <- mean(difFinal > 0)
## Probabilidade do Treze ser campeão
probs[2] <- mean(difFinal < 0)

names(probs) <- c("Campinense", "Treze")

probs*100

barplot(probs*100, ylim=c(0,100))
```

AGRADECIMENTOS

Agradeço primeiramente a Deus por tudo que me concedeu, por todas as bênçãos derramadas em minha vida e por ter permitido com a sua graça chegar até aqui, mesmo com todas as dificuldades encontradas no decorrer do curso.

À minha mãe, ao meu irmão e a todos da minha família que me motivaram, incentivaram e me ajudaram só tenho a agradecer, sem eles eu não conseguiria, meu muito obrigado e que Deus possa abençoá-los e protegê-los sempre.

Ao Professor Gustavo por ter aceito o convite de me conduzir neste trabalho de pesquisa, por toda a ajuda, ensinamentos e paciência comigo, muito obrigado por tudo, sem o seu auxílio neste trabalho eu não conseguiria realizá-lo, muito obrigado só tenho a agradecer, que DEUS continue abençoando a sua vida.

Agradeço à Universidade Estadual da Paraíba e a todos os professores e professoras do curso de Estatística pelos elevados e qualificados ensinamentos durante o curso.