UNIVERSIDADE ESTADUAL DA PARAÍBA

CENTRO DE CIÊNCIAS E TECNOLOGIA

DEPARTAMENTO DE ESTATÍSTICA

PEDRO AUGUSTO FERREIRA DA SILVA

CÁLCULO DE PROBABILIDADES NO FUTEBOL: UMA APLICAÇÃO NO CAMPEONATO BRASILEIRO DE 2019

CAMPINA GRANDE - PB

2019

PEDRO AUGUSTO FERREIRA DA SILVA

CÁLCULO DE PROBABILIDADES NO FUTEBOL: UMA APLICAÇÃO NO CAMPEONATO BRASILEIRO DE 2019

Trabalho de Conclusão de Curso apresentado ao curso de Bacharelado em Estatística do Departamento de Estatística do Centro de Ciências e Tecnologia da Universidade Estadual da Paraíba em cumprimento às exigências legais para obtenção do título de bacharel em Estatística.

Orientador: Prof. Gustavo Henrique Esteves

CAMPINA GRANDE - PB

2019

S586c Silva, Pedro Augusto Ferreira da.

Cálculo de probabilidades no futebol [manuscrito] : uma aplicação no Campeonato Brasileiro de 2019 / Pedro Augusto Ferreira da Silva. - 2019.

33 p. : il. colorido.

Digitado.

Trabalho de Conclusão de Curso (Graduação em Estatística) - Universidade Estadual da Paraíba, Centro de Ciências e Tecnologia , 2019.

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

1. Probabilidade. 2. Métodos de estimação. 3. Futebol. I. Título

21. ed. CDD 519.2

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

Pedro Augusto Ferreira da Silva

Cálculo de Probabilidades no Futebol: Uma Aplicação no Campeonato Brasileiro de 2019

Trabalho de Conclusão de Curso apresentado ao curso de Bacharelado em Estatística do Departamento de Estatística do Centro de Ciências e Tecnologia da Universidade Estadual da Paraíba em cumprimento às exigências legais para obtenção do título de bacharel em Estatística.

Trabalho aprovado em 05 de dezembro de 2019.

BANCA EXAMINADORA

Prof. Gustavo Henrique Esteves
Universidade Estadual da Paraíba

Prof. Dr. Tiago Almeida de Oliveira
Universidade Estadual da Paraíba

Prof. Dr. Sílvio Fernando Alves Xavier Júnior
Universidade Estadual da Paraíba

Dedico este trabalho primeiramente a Deus que permitiu que tudo pudesse ser realizado. À minha mãe Simone, pelo apoio incondicional em todos os momentos. À minha tia e madrinha, Cristiane, que sempre foi a minha maior incentivadora nos meus estudos. Aos meus familiares, que me apoiaram de diversas maneiras nessa etapa tão importante da minha vida. A todos os meus amigos por me apoiarem de todas as formas neste sonho. E ao meu orientador Gustavo pelos valiosos ensinamentos e entusiasmo com a pesquisa.

Agradecimentos

Ser estatístico se tornou um sonho que floresceu durante esta minha longa jornada. Com desafios e muita luta, gostaria muito de agradecer primeiramente a Deus por ter guiado essa minha trajetória na graduação.

Em primeiro lugar, gostaria muito de agradecer a minha Tia Cristiane. Desde os tempos de colégio, você sempre confiou no meu potencial e hoje estou aqui, terminando esta trajetória, graças especialmente a você. Muitíssimo obrigado!

Gostaria também de agradecer em especial a minha mãe Simone, por ter sido uma das minhas maiores inspirações durante esta minha jornada. Sem dúvidas, todo o seu esforço valerá a pena minha mãe. Muitíssimo obrigado!

Não posso deixar também de agradecer as minhas tias Mônica, Márcia, Cíntia e Jane, além dos meus tios Aluízio, Lenílton, Marcelo, Múcio e Alexandre por terem apoiado de alguma forma esta minha jornada. Também agradeço ao meu querido irmão Pablo e aos meus primos Caio e Júlio. Muito obrigado!

Não só a minha família foi um combustível para que eu pudesse chegar até aqui. Outro pilar fundamental foram meus grandes amigos que construí durante esta jornada. E quero agradecer especialmente cada um de vocês: Adrielle, Débora, Hiago, Iago, Leo, Lucas, Adenilson, Janaína, Gustavo e outros tantos que levarei pelo resto da minha vida. Sem vocês eu não conseguiria chegar até aqui. Vocês foram uma das minhas inspirações. Muitíssimo obrigado, meus amigos!

E por fim, não mais do que importante, gostaria de agradecer a todos os meus professores da graduação. Desde o primeiro período até então. Obrigado por vocês passarem um pouco de seus conhecimentos. Vocês, sem dúvidas, foram uma das grandes motivações por ter me apaixonado por este curso tão fantástico que é a Estatística. Muito obrigado!

"A resposta certa não importa nada: o essencial é que as perguntas estejam certas."
(Mário Quintana)

Resumo

O presente trabalho teve como objetivo abordar alguns modelos de previsão em resultados de partidas de futebol, por meio de simulação, para o Campeonato Brasileiro de 2019, com o intuito de calcular a probabilidade de título dos times participantes em vencerem a competição, bem como utilizar análise de agrupamento para obter conclusões sobre times que se destacaram na competição. Na área futebolística, a Probabilidade e a Estatística são ferramentas que vêm ganhando muita importância para previsões de partidas, além de mostrar informações sobre os clubes de futebol aos torcedores. Neste sentido, os métodos de estimação Soma e Diferença e Log-Lineares são um dos métodos aplicados para tais previsões de maneira a se utilizar nas estimativas dos parâmetros associados aos métodos. No início do campeonato o Palmeiras e o Flamengo possuíam probabilidades consideradas de se consagrarem campeões. A cada rodada realizada, o Flamengo se destacou devido a diversos efeitos como a mudança de treinador durante o campeonato, consagrando o título da competição. Além disso, a grande disparidade entre esses dois times resultou na análise de agrupamento com a formação de 3 grupos, em que Flamengo e Palmeiras se destacaram dos demais clubes da competição, uma vez que tais equipes possuem um ótimo elenco e um excelente poder financeiro.

Palavras-chaves: Futebol. Probabilidade. Métodos de Estimação.

Abstract

The objective of this study was to approach some prediction models in soccer match results, through simulation for the 2019 Brazilian Championship, in order to calculate the probability of the participating teams to win the competition, as well as to use cluster analysis in order to get to a conclusions about teams that stood out. When talking about soccer, Probability and Statistics have been tools that are gaining much importance for match predictions. Also providing important information about soccer clubs to its fans. In this sense, the Sum and Difference and Log-Linear estimation methods are one of the methods applied for such predictions, in order to use for parameters estimation associated with the methods. At the beginning of the championship, Palmeiras and Flamengo were considered likely to be champions. For each round, Flamengo stood out due to the different aspects such as the change of coach during the championship, resulting the team to earn the title of the competition. In addition, the huge difference between these two teams resulted in the 3 grouping analysis, in which Flamengo and Palmeiras stood out from the other clubs of the competition, as such teams have a great group and excellent financial power.

Key-words: Soccer. Probability. Estimation Methods.

Lista de ilustrações

Lista de tabelas

Sumário

1 Introdução

Com milhões de torcedores e times ao redor do mundo, o Futebol proporciona aos adeptos uma mistura de emoções. A prática futebolística entre milhões de pessoas, ora como torcedor de um time de coração, ora como um praticante deste desporto, faz com que este seja o esporte mais popular do mundo. A Estatística, como ciência, dá um suporte de informações e análises quantitativas e qualitativas sobre diversas temáticas no mundo, inclusive no futebol por meio de dados sobre histórico de confrontos de times, número médio de cartões em que um árbitro aplica durante uma partida, ou até mesmo a probabilidade de um time ser campeão em um determinado campeonato.

No Brasil, o futebol é o esporte mais popular do país. Com vastos clubes espalhados em todo lugar do território a quantidade de campeonatos, promovidos pela Confederação Brasileira de Futebol, fez com que este esporte se popularizasse mais no âmbito nacional. Por sua vez, o Campeonato Brasileiro iniciado em 1959, ainda chamado como Taça Brasil, foi o primeiro torneio nacional de grande impacto entre os torcedores, tendo como o primeiro campeão o Esporte Clube Bahia. Em 1971, houve uma reestruturação e passou-se chamar Campeonato Brasileiro, nome que permanece até hoje. Nestes 60 anos de Campeonato, o grande número de clubes campeões nacionais fez com que o Campeonato fosse um dos mais disputados do mundo, e tornou-se uma grande vitrine para os torcedores apaixonados pelos seus clubes de coração.

Ao passar dos anos o Campeonato Brasileiro torna-se cada vez mais disputado entre os clubes com diversos objetivos como vagas para torneios internacionais e permanência na primeira divisão, mas sobretudo almejando sempre o título. No ano de 2019, vinte clubes de quatro regiões do país disputaram entre turno e returno durante todo o ano, totalizando 380 jogos durante o campeonato. Nos últimos anos, a Série A do Campeonato Brasileiro vem se modernizando entre os clubes e pela CBF. No âmbito dos times, o papel do analista de desempenho tornou-se fundamental para a evolução dos atletas. Com diversas informações de cada jogador, este profissional proporcionou uma vasta quantidade de dados sobre partidas e jogadores não só do clube, mas também de outros clubes. Já no âmbito da instituição, a inclusão de tecnologias sobretudo na arbitragem proporcionou uma grande profissionalização do futebol nacional. Tal incremento possibilita ao torcedor um campeonato de qualidade, colocando a Série A como um dos principais campeonatos nacionais do mundo.

Atualmente, com o avanço da tecnologia, as estatísticas no campo futebolístico proporcionaram ao torcedor uma infinidade de informações no que se concerne à previsão de resultados das partidas entre times. Segundo Suzuki (2007), a grande procura de

estatísticos nos países proporcionou a criação de diversos modelos com o objetivo de prever os resultados de partidas no futebol. Além disso, Suzuki (2007) diz que diversos fatores podem influenciar em uma partida, haja vista que tais fatores podem se tornar variáveis nesses modelos estatísticos. Para esta modelagem, considera-se um modelo probabilístico, de maneira a se obter as estimativas dos parâmetros conforme o modelo adequado.

Em uma partida de futebol existem três possibilidades em que o jogo pode terminar, de modo que ou time mandante pode sair vitorioso, ou o visitante triunfar ou a partida terminar empatada. Dessa maneira, esses resultados dependerão do número de gols que cada time marque durante este evento. Na teoria de Probabilidades, a distribuição apropriada para este tipo de evento é a *Poisson*, uma vez que a partida de futebol é um processo de contagem, do qual os times mandante e visitante buscam marcar gols durante o jogo. Se o time mandante for analisado, o número de gols que este poderá fazer durante a partida pode seguir uma distribuição de *Poisson* univariada. Por outro lado, se for observado o número de gols do time visitante, este também seguirá uma distribuição de *Poisson* univaridada. Neste sentido, surge a ideia da aplicabilidade da distribuição de *Poisson* Bivariada, pois nos modelos probabilísticos em previsão de resultados, o interesse está em se modelar as probabilidades de um time ganhar, perder ou empatar a partida de interesse.

De acordo com AlMuhayfith, Alzaid e Omair (2016), os modelos de contagem bivariados são utilizados quando existem duas variáveis em processos de contagem correlacionados e que estas precisam ser estimadas de maneira conjunta. Por exemplo, em uma loja de eletrodomésticos o número de clientes que se dirigem ao caixa para realizar uma troca por defeito ou por arrependimento é um processo de contagem bivariado. Neste sentido, a distribuição de *Poisson* bivariada proposta por Holgate (1964) e apresentada também por Johnson, Kemp e Kotz (2005) é um dos modelos mais utilizados para processos de contagem bivariados. Existem diversos modelos de *Poisson* bivariados. Entretanto, a *Poisson* bivariada de "Holgate" é a mais usual em resultados de partidas de futebol. Dessa forma, a utilização de modelos estatísticos utilizando esta distribuição tornou-se bastante fundamental para a literatura.

Neste sentido, o presente trabalho abordará os modelos de previsão em partidas de futebol, aplicando tais modelos com simulação no Campeonato Brasileiro de 2019 em cada rodada para prever qual o eventual Campeão Brasileiro deste ano. Além disso, objetivou-se também, por meio de técnicas multivariadas, comparar diversos comportamentos entre os times que participaram do campeonato, assim como aqueles que se destacaram durante a competição.

2 Fundamentação Teórica

2.1 Classe de *Poisson* Bivariada de Holgate

Na literatura existem diversos modelos probabilísticos para a previsão de resultados em partidas de futebol. Segundo Arruda (2000), a distribuição *Poisson* Bivariada "de Holgate" é a mais apropriada, pois ela possui uma propriedade de ser infinitamente divisível, ou seja, as distribuições conjuntas e a marginal têm a mesma distribuição de uma soma de n vetores aleatórios independentes, identicamente distribuídos. Dessa forma, a utilização de modelos estatísticos utilizando esta distribuição, tornou-se bastante fundamental para a literatura. De acordo com Suzuki (2007), tal propriedade se dá pelo fato de que uma partida de futebol é ocorrida em um intervalo contínuo de tempo e não em um valor pré-estabelecido, como uma partida de Tênis ou Vôlei, por exemplo.

A classe de distribuição *Poisson* bivariada de Holgate (1964) é definida como uma distribuição conjunta das variáveis $X = S_1 + S_{12}$ e $Y = S_2 + S_{12}$ em que S_1, S_2, S_{12} são variáveis aleatórias com distribuição Poisson univaridada. Dessa forma, a distribuição *Poisson* bivariada é definida como três processos de *Poisson* independentes, no qual S_1, S_2 e S_12 são os números de observações para o mandante, visitante e a conjunta em cada processo do determinado período de tempo, respectivamente. Estas variáveis possuem taxas de ocorrência λ_1, λ_2 e λ_{12}, ao qual as variáveis X e Y têm distribuição conjunta *Poisson* Bivariada.

A seguir serão apresentados os métodos de estimação de parâmetros para as previsões de resultados em partidas de futebol. Cabe ressaltar que para todos os métodos citados abaixo, (X, Y) representa uma variável aleatória bivariada com distribuição *Poisson* de Holgate, onde X e Y denotam, respectivamente, os números de gols marcados pelos times mandante e visitante em uma eventual partida de futebol.

2.2 Métodos Soma e Diferença

Estes métodos, propostos por Arruda (2000) e adaptados por Suzuki (2007), têm como base as seguintes propriedades apresentadas nas equações a seguir:

$$E(X - Y) = \lambda_1 - \lambda_2$$

$$E(X + Y) = \lambda_1 + \lambda_2 + 2\lambda_{12}.$$

Desse modo, tais métodos se diferem pelo fato de que o método Soma e Diferença(ou SD) 0 considera o parâmetro de covariância nulo, isto é, $\lambda_{12} = 0$, enquanto o método SD

1 considera tal parâmetro não nulo, sendo necessária sua estimação. Vale salientar que mais adiante serão apresentados os resultados apenas para o método SD 1, uma vez que a covariância pode ser estimada neste modelo.

2.2.1 Método Soma e Diferença 0

Em primeiro lugar, neste método o parâmetro da covariância é considerado nulo, ou seja $\lambda_{12} = 0$, não sendo necessária sua estimação. Diante disso, as variáveis referentes ao número de gols marcados entre as equipes mandante e visitante em cada partida são independentes entre si. Por sua vez, a covariância entre as variáveis será abordada no método SD 1. Os parâmetros de interesse λ_1 e λ_2 são obtidos no sistema de equações abaixo, de acordo com as propriedades da distribuição de Holgate (ARRUDA, 2000):

$$\begin{cases} E[X-Y] = \lambda_1 - \lambda_2 \\ E[X+Y] = \lambda_1 + \lambda_2 \end{cases}$$

e os estimadores para λ_1 e λ_2 dados por

$$\begin{cases} \hat{\lambda}_1 = \frac{\hat{E}[X-Y]+\hat{E}[X+Y]}{2} \\ \hat{\lambda}_2 = \frac{\hat{E}[X+Y]-\hat{E}[X-Y]}{2} \end{cases}$$

de modo que $E[X+Y]$ e $E[X-Y]$ são estimados através de modelos lineares definidos por

$$(X+Y)_i = \mathbf{G_i}\alpha + \epsilon_{ai} \tag{2.1}$$

e

$$(X-Y)_i = \mathbf{H_i}\beta + \epsilon_{bi} \tag{2.2}$$

onde $i = 1, 2, 3, ..., n; n$ é o número de jogos no banco de dados e ϵ_{ai} e ϵ_{bi} são erros independentes com médias iguais a 0.

Além disso, a Equação (2.1) modela o total de gols marcados entre as equipes no i-ésimo jogo, é um vetor associado aos times participantes do campeonato mais um termo referente ao local, constituído por $N+1$ parâmetros no qual N é o número de times do banco de dados. Cada linha da matriz $\mathbf{G}$ refere-se aos *status* dos times participantes para uma dada partida mais um componente destinado ao fator local (existência ou não de fator campo). Este *status* é uma variável de incidência em que ela assume o valor 1 se a equipe participa do i-ésimo jogo ou 0 caso contrário. De acordo com Suzuki (2007), esta atribuição se dá pelo fato de que há uma independência na identificação dos times participantes, ou seja, o time ser mandante ou visitante não interfere no resultado de $(X+Y)_i$. Por sua vez, a variável Local também assume valores 1 se há algum fator referente ao time mandante, ou seja, jogar em sua casa, e 0 caso o time mandante não jogue na sua casa por outro fator, seja ele ter vendido o mando de campo, ou por punição de mando de campo.

Outrossim, Suzuki (2007) diz que ambos os métodos da família SD são oriundos através de modelos lineares sem intercepto, haja vista que uma partida de futebol sempre começa pelo placar de 0×0 e nunca com algum placar em vantagem para um dos clubes. A título de exemplificação, a organização da matriz $\mathbf{G}$ será apresentada na seção 2.4.

Por outro lado, a Equação (2.2) modela a diferença de gols marcados entre os times, ou seja, é o número de gols marcados pelo mandante menos o número de gols marcados pelo visitante no i-ésimo jogo em questão. Aqui, o vetor é formado por $N + 1$ parâmetros no qual N refere-se aos clubes do banco de dados mais o parâmetro referente ao fator local. A matriz $\mathbf{H}$, por sua vez, é atribuída ao status entre as equipes participantes do jogo observado, além de uma componente referente ao local da partida. Vale salientar que diferente da matriz $\mathbf{G}$, os valores referentes a variável status na matriz $\mathbf{H}$ são definidos como 1 se o time é mandante da partida, -1 se é visitante ou 0 se este não participa do jogo em questão. Esta distinção se dá pelo fato de que neste modelo linear há diferença entre os resultados da partida, isto é, 1×0 é diferente de 0×1 aqui. Por sua vez, a componente do local é uma variável que atribui o valor 1 se o mandante joga em seus domínios e 0 se joga em campo neutro.

2.2.2 Método Soma e Diferença 1

O método SD 1 inclui a estimativa da covariância entre as variáveis de interesse, denotada por λ_{12}. De acordo com as propriedades das variâncias e covariância da distribuição de Holgate, demonstradas em Arruda (2000), tem-se que

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

Desse modo, as expressões obtidas para os valores dos parâmetros λ_1,λ_2 e λ_{12} são explicitadas no sistema de equações abaixo:

$$\begin{cases} E(X-Y) & = & \lambda_1 - \lambda_2 \\ E(X+Y) & = & \lambda_1 + \lambda_2 + 2\lambda_{12} \\ E[(X+Y)^2] - [E(X+Y)]^2 & = & \lambda_1 + 4\lambda_{12} + \lambda_2 \end{cases}$$

de modo que,

$$\begin{cases} \lambda_1 = \frac{E(X-Y)+2E(X+Y)-\left\{E[(X+Y)^2]-[E(X+Y)]^2\right\}}{2} \\ \lambda_2 = \frac{2E(X+Y)-E(X-Y)-\left\{E[(X+Y)^2]-[E(X+Y)]^2\right\}}{2} \\ \lambda_{12} = \frac{\left\{E[(X+Y)^2]-[E(X+Y)]^2\right\}-E(X+Y)}{2} \end{cases}$$

e, assim, as estimativas dos parâmetros

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

A estimação da expressão $[E(X+Y)]^2$ se dá através do modelo linear

$$\left[(X+Y)^2\right]_i = \mathbf{G_i}\gamma + \epsilon_{ci} \tag{2.3}$$

onde $i = 1, 2, 3, ..., n; n$ é o número de partidas contidas no banco de dados e ϵ_{ci} é o erro independente com média igual a 0. Os modelos lineares (2.1) e (2.2) são utilizados para estimar $E(X+Y)$ e $E(X-Y)$, respectivamente.

Por sua vez, o modelo (2.3) refere-se ao quadrado do número de gols marcados na i-ésima partida da amostra e o vetor é constituído por $N+1$ parâmetros, novamente onde N é o parâmetro referente ao número de clubes mais uma componente destinada ao local. Ademais, a matriz $\mathbf{G}$ é idêntica àquela construída para o método SD 0, descrito anteriormente.

2.3 Métodos Log-Lineares

Os métodos log-Lineares, ou Chance, propostos por Lee (1997), na revista de próprio nome, são métodos que têm como finalidade estimar parâmetros de interesse através da média de gols marcados pelos clubes que disputam a partida. Por outro lado, tais médias constituem a força dos times no que se concerne a qualidade do seu adversário e a vantagem do time jogar em seu domínio.

2.3.1 Método Log-Linear 1

O método log-Linear 1, abordado em Arruda (2000), não considera o parâmetro da covariância, de maneira análoga ao método SD 0. Desse modo, não importa a quantidade de gols marcados pelo time da casa. A distribuição para o time visitante será a mesma. Dessa maneira, $E(X) = \lambda_1$ e $E(Y) = \lambda_2$ serão estimados através de um modelo log-Linear de *Poisson*. Cabe ressaltar que para esta modelagem a notação muda, onde os números de gols marcados pelos times mandante e visitante no i-ésimo jogo serão agora denotados por X_{2i-1} e X_{2i}, respectivamente.

Definição 2.1. Seja X_{2i-1} e X_{2i} o número de gols marcados pelas equipes mandante e visitante. Logo, $X_{2i-1} \sim Poisson\{\lambda_{2i-1}\}$ e $X_{2i} \sim Poisson\{\lambda_{2i}\}$ independentes entre si. —

A função de distribuição de probabilidade para X_{2i-1} é dada por:

$$f(x_{2i-1}) = \frac{e^{\lambda_{2i-1}}\lambda_{2i-1}^{x_{2i-1}}}{x_{2i-1}!}$$

e a função de distribuição de probabilidade para X_{2i} é

$$f(x_{2i}) = \frac{e^{\lambda_{2i}}\lambda_{2i}^{x_{2i}}}{x_{2i}!}.$$

Por sua vez, o logaritmos das funções de verossimilhança são escritos da seguinte maneira

$$\begin{cases} l(\lambda_{2i-1}, X_{2i}) = -\lambda_{2i-1} + x_{2i-1}\log\lambda_{2i-1} - \log(x_{2i-1})! \\ l(\lambda_{2i}, X_{2i}) = -\lambda_{2i} + x_{2i}\log\lambda_{2i} - \log(x_{2i})!. \end{cases} \tag{2.4}$$

Segundo Suzuki (2007), a função de ligação para o modelo log-linear de Poisson é dada por

$$\lambda_j = e^{\mathbf{U_j}\beta} \tag{2.5}$$

para $j = 1, 2, ..., 2n$, $\mathbf{U_1}, \mathbf{U_2}, ..., \mathbf{U_{2n}}$ são vetores de covariáveis e β um vetor de parâmetros. Dessa forma, substituindo (2.5) em (2.4), tem-se que

$$\begin{cases} l(\lambda_{2i-1}, X_{2i}) = -e^{\mathbf{U_{2i-1}}B} + x_{2i-1}\mathbf{U_{2i-1}}\beta - \log(x_{2i-1})! \\ l(\lambda_{2i}, X_{2i}) = -e^{\mathbf{U_{2i}}B} + x_{2i}\mathbf{U_{2i}}\beta - \log(x_{2i})!. \end{cases}$$

Assim, dado a realização de n partidas, o modelo é definido por

$$l(\lambda_1, \lambda_2, ..., \lambda_{2n-1}, \lambda_{2n}, X_1, X_2, ..., X_{2k-1}, X_{2k}) = \sum_{j=1}^{2n}\left(-e^{\mathbf{U_j}\beta} + x_j\mathbf{U_j}\beta - \log(x_j!)\right)$$

de maneira que x_{2i-1} e x_{2i} refere-se ao número de gols marcados pelas equipes mandante e visitante, respectivamente, no i-ésimo jogo em questão. O vetor β é composto por $2N + 2$ parâmetros, onde um parâmetro é destinado ao intercepto, dois parâmetros destinados ao ataque e a defesa de cada time composto no banco de dados e um parâmetro destinado ao local onde se realiza a partida. Por sua vez, as matrizes-linha $\mathbf{U_{2i-1}}$ e $\mathbf{U_{2i}}$ são compostas de $2N + 2$ componentes, onde a primeira entrada sempre será 1 (destinada ao intercepto) e as demais componentes referente ao *status* de cada equipe. Nesta matriz, as primeiras entradas após do intercepto são destinadas aos fatores de ataque para cada equipe participante do jogo, seguida dos fatores de defesa, além do local.

Desse modo, os valores destinados ao *status* na matriz $\mathbf{U_{2i-1}}$ será 1 se a equipe mandante participa do jogo em questão e 0 caso contrário para o ataque. Em relação a defesa, receberá -1 se a equipe participante da partida em questão for visitante e 0 caso contrário. De maneira análoga, as componentes referente aos *status* das equipes na matriz $\mathbf{U_{2i}}$ será 1 se a equipe visitante participa do jogo em questão e 0 caso contrário para o ataque. Além disso, para a defesa, receberá -1 se a equipe for mandante da partida e 0 caso contrário. Por fim, no que se concerne ao local na matriz $\mathbf{U_{2i-1}}$, receberá 1 se o mandante joga em seus domínios e 0 caso contrário, e, analogamente, na matriz $\mathbf{U_{2i}}$ receberá 1 se a equipe visitante joga no local da partida e 0 caso contrário.

2.3.2 Método Log-Linear 2

Proposto por Arruda (2000), este método é semelhante ao log-Linear 1, no entanto, inclui a estimação da covariância entre X e Y. Conforme a propriedade da distribuição

de "Holgate" em que as esperanças marginais de X e Y são dadas por $E(X) = \lambda_1 + \lambda_2$ e $E(Y) = \lambda_2 + \lambda_{12}$, respectivamente. O objetivo deste método é resultar em uma linearização de maneira tal que λ_{12} possa ser decomposta em uma parcela comum, assim como uma parcela (λ_1 ou λ_2) seja relacionada com as distribuições de X e Y. Neste sentido, uma linearização para o modelo (2.5) é definida como:

$$X_i = \beta_0 + \mathbf{U_i}'\beta' + \epsilon_{bi}$$

ou, ainda,

$$E(X_i) = \beta_0 + \mathbf{U_i}'\beta', \tag{2.6}$$

em que ϵ_{bi} são erros independentes com médias iguais a 0, $\mathbf{U_i}$ é a matriz $\mathbf{U}$ e β' é o mesmo vetor de β definidos no método log-Linear 1, com a exceção da coluna de $\mathbf{U}$ e da componente de β destinadas ao intercepto.

De acordo com Suzuki (2007), é possível atribuir simetrias entre a parcela comum e o processo comum, assim como as parcelas específicas e os processos específicos das variáveis X e Y, conforme à construção da distribuição de "Holgate" através de processos de Poisson. Neste sentido, designando um paralelismo entre $E(X) = \lambda_1 + \lambda_{12}$ e o modelo (2.6) tem-se que

$$E(X) = \lambda_{12} + \lambda_1 = \beta_0 + \mathbf{U_i}'\beta', \tag{2.7}$$

onde λ_{12} e β_0 são as parcelas comuns e λ_1 e $\mathbf{U_i}'\beta'$ são as parcelas específicas. Então, pode-se observar por paralelismo que há uma correspondência entre os termos acima, podendo concluir que $\lambda_{12} = \beta_0$. Desse modo, as expressões para os parâmetros de interesse λ_1, λ_2 e λ_{12} é definido no sistema de equações:

$$\begin{cases} E(X) = \lambda_1 + \lambda_{12} \\ E(Y) = \lambda_2 + \lambda_{12} \\ \beta_0 = \lambda_{12} \end{cases}$$

e, assim, obtendo os seguintes estimadores

$$\begin{cases} \hat{\lambda_1} = \hat{E}(X) - \hat{\beta}_0 \\ \hat{\lambda_2} = \hat{E}(Y) - \hat{\beta}_0 \\ \hat{\beta}_0 = \hat{\lambda_{12}} \end{cases}$$

portanto, $E(X)$, $E(Y)$ e β_0 são estimados pelo modelo linear

$$X_j = \mathbf{U_j}\beta + \epsilon_{bi},$$

para $j = 1, 2, ..., 2n$ e ϵ_{bi} são erros com distribuição normal independentes com médias iguais a 0. Neste modelo, X_j, β e $\mathbf{U_j}$ têm a mesma equivalência do método log-linear 1.

Uma observação neste método é que as estimativas dos parâmetros para os times envolvidos em uma partida em particular podem resultar valores negativos, o que ocasionaria um problema nesta modelagem. Para contornar tal situação, Arruda (2000) sugere que se realize uma projeção do ponto estimado para o ponto mais próximo que esteja dentro dos valores possíveis para os parâmetros, isso equivale a projetar os valores a zero. Isto ocasionaria outro problema, pois se, por exemplo, $\lambda_1 = 0$, o que implica $S_1 \equiv 0$ e, sequencialmente, $X = S_{12}$ e $Y = S_2 + S_{12}$. Neste sentido, como S_2 e S_{12} são positivos, esta relação implicaria que o time visitante sempre terá uma probabilidade maior que o time mandante. Para contornar esta situação, Arruda (2000) sugere que se adicione um valor pré-estabelecido, no lugar de zerar a estimativa do parâmetro.

2.4 Exemplo

Para ilustrar a construção das matrizes dos métodos Soma e Diferença descritos acima, construiu-se um torneio hipotético entre os times Campinense-PB, Treze-PB, Botafogo-PB e Atlético-PB. Neste Torneio, cada time jogou apenas uma vez contra cada um dos adversários com os mandos de campo definidos ao acaso. Os resultados obtidos para todas as seis partidas hipotéticas, bem como os locais no qual os jogos teriam acontecido, estão listados a seguir:

Tabela 1 – Tabela referente ao torneio hipotético entre os clubes Atlético-PB, Botafogo-PB, Campinense-PB e Treze-PB.

Partidas Realizadas no Torneio					
Mandante		x		Visitante	Local
Treze-PB	1	x	0	Campinense-PB	Amigão
Botafogo-PB	1	x	2	Atlético-PB	Amigão
Atlético-PB	2	x	3	Treze-PB	Cajazeiras
Campinense-PB	3	x	2	Botafogo-PB	Amigão
Treze-PB	3	x	0	Botafogo-PB	Amigão
Campinense-PB	2	x	1	Atlético-PB	Amigão

Vale salientar que o primeiro jogo entre Treze-PB e Campinense-PB foi no Amigão, logo o campo é neutro para ambos, uma vez que os dois clubes jogam neste estádio. Além disso, o jogo entre Botafogo-PB e Atlético-PB não foi realizado em João Pessoa, casa do Botafogo, e sim na cidade de Campina Grande, por outros fatores extra-campo. Por fim, cada partida admitirá peso igual, não havendo nenhum acréscimo nos valores das estimativas e dos resultados finais.

Após o término dos jogos, o jogo final deste torneio seria Treze-PB x Campinense-PB no estádio Presidente Vargas, casa do Treze. Para estimar os parâmetros dos times desta partida será utilizado o método SD 1, de acordo com os resultados obtidos nas

partidas do torneio. Inicialmente, a construção da matriz $\mathbf{G}$ é dada como

$$\mathbf{G} = \begin{pmatrix} 0 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 \end{pmatrix}.$$

A primeira linha desta matriz é referente ao primeiro jogo do torneio e assim sucessivamente. Por sua vez, cada coluna refere-se aos times do torneio seguindo a ordem alfabética e a última coluna refere-se ao local da partida. Já o vetor α é dado em ordem alfabética dos clubes, ou seja,

$$\alpha = \begin{pmatrix} \alpha_{Atlético} \\ \alpha_{Botafogo} \\ \alpha_{Campinense} \\ \alpha_{Treze} \\ \alpha_{Local} \end{pmatrix}$$

e o vetor $(X + Y)$ é dado por

$$(X + Y) = \begin{pmatrix} 1 \\ 3 \\ 5 \\ 5 \\ 3 \\ 3 \end{pmatrix},$$

em que cada informação é referido ao total de gols marcados por ambos os times na partida. Por outro lado, a matriz $\mathbf{H}$ será definida da seguinte maneira:

$$\mathbf{H} = \begin{pmatrix} 0 & 0 & -1 & 1 & 0 \\ -1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & -1 & 1 \\ 0 & -1 & 1 & 0 & 1 \\ 0 & -1 & 0 & 1 & 1 \\ -1 & 0 & 1 & 0 & 1 \end{pmatrix},$$

onde valores negativos contidos na matriz acima refere-se aos times visitantes da partida. Analogamente ao vetor α, o vetores β e γ podem ser preenchidos com a representação de

cada clube, além do local. Por sua vez, $(X - Y)$ é preenchido da seguinte forma:

$$(X - Y) = \begin{pmatrix} 1 \\ -1 \\ -1 \\ 1 \\ 3 \\ 1 \end{pmatrix}$$

e, por fim, o vetor $(X + Y)^2$ é o quadrado dos valores obtidos em $(X + Y)$.

Os estimadores λ_1, λ_2e λ_{12} em relação ao jogo final são obtidos por meio de

$$\hat{E}(X + Y) = \mathbf{G_7}\hat{\alpha}$$

$$\hat{E}(X - Y) = \mathbf{H_7}\hat{\beta}$$

$$\hat{E}[(X + Y)^2] = \mathbf{G_7}\hat{\gamma}$$

onde $\hat{\alpha}$,$\hat{\beta}$,$\hat{\gamma}$ são estimados através de mínimos quadrados entre as soma dos quadrados da diferenças $(X + Y)_i - \mathbf{G_i}\hat{\alpha}$, $(X - Y)_i - \mathbf{H_i}\hat{\beta}$ e $[(X + Y)^2]_i - \mathbf{G_i}\hat{\gamma}$. De acordo com os cálculos obtidos através do *software* R (R Core Team, 2019), as estimativas obtidas para os parâmetros dos clubes finalistas do torneio são $\hat{\lambda}_{Treze} = 1,875$, $\hat{\lambda}_{Campinense} = 0,125$ e $\hat{\lambda}_{12} = 0,5$. Por fim, as probabilidades obtidas paras os possíveis resultados, são:

- Probabilidade de vitória do Treze-PB calculada em 0,8107 (equivalente a 81,07%);
- Probabilidade de empate calculado em 0,1690 (16,9%);
- Probabilidade de vitória do Campinense-PB calculada em 0,0202 (2,02%).

2.5 Simulação

No campo da Estatística, a simulação é uma ferramenta de suma importância em diversas aplicações. Por meio de técnicas computacionais, a simulação fornece para o usuário resultados oriundos de modelos estatísticos, com o objetivo de se obter resultados aproximados. A ideia de simulação é gerar um grande número de repetições de uma variável aleatória a partir de algum modelo probabilístico de maneira a se obter valores conclusivos de uma proporção total obtida.

No Campeonato Brasileiro, a aplicação do método SD1 é necessária para calcular a probabilidade do time ganhar uma determinada partida ou rodada da competição. Como o campeonato possui 38 rodadas e 10 jogos por rodada, a simulação é fundamental para simular todos os 380 jogos do Campeonato a partir das estimativas obtidas pelo método SD1. Neste sentido, a probabilidade do clube ser campeão é calculada por meio de

simulação. Através dos 380 jogos a serem realizados, bem como um banco de dados contendo o histórico de partidas dos 20 clubes participantes deste ano, a simulação é realizada por um número pré-determinado pelo usuário. Este número deverá ser significantemente grande para que se tenha maior confiança nos resultados obtidos. Desta forma, um número considerável é de dez mil repetições nestas partidas a serem realizadas no campeonato.

Por exemplo, se o usuário deseja obter uma probabilidade estimada do time ser campeão com dez mil simulações, computacionalmente o campeonato será replicado dez mil vezes por completo de maneira que cada repetição é armazenada em um objeto contando quantas vezes o determinado time foi campeão. A probabilidade do clube ganhar o campeonato é dada pela proporção do número de vezes que este foi campeão nas simulações dividido pelo número total de simulações realizadas, como apresentado na expressão a seguir.

$$\hat{p} = \frac{n_j}{n_s}$$

onde, $\hat{p}$ é a probabilidade estimada do time ser campeão, n_j é o número de vezes em que o time foi campeão e n_s é o número total de simulações.

2.6 Técnicas Multivariadas

Uma das principais áreas em Estatística no qual se trata em trabalhar com diversas variáveis conjuntamente é o ramo da Estatística Multivariada. Tal ramo se caracteriza por diversas técnicas multivariadas no qual se objetiva em inferir sobre a realidade. Na Estatística Multivariada existem diversas técnicas, dentre as quais a Análise de Agrupamento (*Cluster*) é uma das principais no que se concerne em agrupar variáveis de maneira mais similar possível(JONHSON; WICHERN, 1998).

Conforme Vicini (2005), a análise de agrupamento consiste em particionar variáveis por meio de uma estrutura homogênea, seguindo algum critério de homogeneidade. Além disso, Vicini (2005) diz que a utilização de várias metodologias é fundamental em análise de agrupamento de maneira que se compare diversos modelos afim de se obter a técnica mais adequada. Para tais modelos utiliza-se uma medida de similaridade com o intuito de se obter os grupos mais similares possíveis.

2.6.1 Medidas de Similaridade e Dissimilaridade

As medidas de Similaridade e Dissimilaridade objetivam identificar os objetos quanto a sua proximidade. Segundo Bussab, Miazaki e Andrade (1990), na medida de similaridade, quanto maior o valor observado menos parecido (mais dissimilares) serão os objetos. Para prosseguir com a análise necessita-se que inicialmente seja escolhida a medida de distâncias a ser utilizada. Estas são calculadas a partir da matriz de similaridade que é utilizada para a obtenção dos grupos. Dessa maneira, pode-se construir uma medida

a partir da outra e vice-versa. Um conceito fundamental na utilização das técnicas de análise de agrupamento é a escolha de um critério que meça a distância entre dois objetos, ou que quantifique o quanto eles são parecidos. Tais critérios utilizados são divididos em duas categorias: medidas de similaridade e de dissimilaridade. A distância euclidiana é um exemplo de dissimilaridade, pois é baseada na diferença entre os objetos. Além disso, a distância euclidiana pode ser definida como a soma da raiz quadrada da diferença entre P e Q em suas respectivas dimensões. No caso n-dimensional, a distância euclidiana entre os pontos $P = (p_1, p_2, ..., p_n)$ e $Q = (q_1, q_2, ..., q_n)$ é definida como:

$$d(P, Q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + ... + (p_n - q_n)^2)} = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}.$$

A utilização da distância euclidiana se aplica melhor a dados não padronizados, ou seja, dados que não tem nenhum tipo de tratamento de adaptação de escala. Uma desvantagem sobre essa medida de distância pode acontecer se houver diferença de escala entre as dimensões, por exemplo, se no eixo X houver a distância em quilômetros, e no eixo Y a mesma estiver em centímetros pensando em termos cartográficos. No momento em que houver a transformação de escala, ou seja, a conversão de centímetros para quilômetros, os resultados euclidianos sofrem uma influência muito grande das dimensões que possuem os maiores valores exercendo influência no resultado final e se não levado em consideração pode tornar a análise de dados enviesada e consequentemente levando a decisões erradas.

3 Aplicação

Com base no ferramental teórico acima, bem como o auxílio do software R (R Core Team, 2019), as análises estatísticas para o Campeonato Brasileiro foram realizadas durante as 34 rodadas até o decorrer deste trabalho, onde o Flamengo se consagrou campeão do campeonato. Utilizou-se o método SD 1 para obter as estimativas dos clubes no torneio. Tais estimativas foram oriundas de resultados do Campeonato Brasileiro de 2018, com o uso dos resultados de todos os 380 jogos realizados, mais a inclusão de alguns jogos de outros campeonatos envolvendo os times que subiram à série A de 2019. No método multivariado, utilizou-se os pacotes `factoextra` (KASSAMBARA; MUNDT, 2017) e `cluster` (MAECHLER et al., 2019) para a análise de agrupamento com a obtenção da matriz e do dendrograma dos times.

Neste sentido, para os times que garantiram o acesso no ano de 2018 para a primeira divisão houve-se a necessidade de buscar partidas entre os clubes participantes e estes que garantiram o acesso. Dois clubes (Avaí-SC e Goiás-GO) que garantiram o acesso haviam jogado a competição anos anteriores, e assim atribuiu-se atribuir ao banco de dados dos jogos do campeonato de 2018. Por outro lado, dois clubes (CSA-AL e Fortaleza) não participavam desta competição há mais de 15 anos. Para tais times, foi necessária a busca por partidas contra times que disputariam o campeonato em demais competições e de diferentes anos. Cabe ressaltar que cada partida atribuiu-se o mesmo peso, não considerando os jogos mais recentes como critério de ponderação.

Além do banco de dados acima, uma planilha contendo as 380 partidas a serem realizadas no Campeonato Brasileiro de 2019 foi utilizada, com o objetivo de simular todos os jogos da competição. Ao término de cada rodada os jogos realizados passavam a integrar o banco de dados, para atualizar as estimativas dos parâmetros para os jogos restantes. A partir daí, todo o processo de simulação de todos os jogos ainda a serem realizados até o final do campeonato foi repetido com o objetivo de se recalcular as estimativas das probabilidades de cada time ser campeão. Para tais simulações, realizou-se dez mil replicações. Todas as tabelas apresentadas neste capítulo estão no Apêndice A.

Antes de iniciar o campeonato, calculou-se as estimativas dos times para a primeira rodada da competição. Tais estimativas tiveram como base para simular as probabilidades estimadas dos times serem campeões antes do campeonato começar. A Tabela 2 apresenta a classificação do campeonato com a probabilidade de cada time ser campeão. Vale ressaltar que todas as colunas estão com valor zero (a menos da probabilidade estimada), pois nenhuma partida do campeonato havia ocorrido.

É importante observar que o Palmeiras-SP possui a maior Probabilidade de ser

Campeão, uma vez que o mesmo é o atual Campeão antes do campeonato se iniciar. Apenas cinco clubes possuem uma probabilidade maior que 0,05 de ser campeão da competição. Com o decorrer do campeonato, o Palmeiras iniciou muito bem a competição, liderando a primeira parte do campeonato. A Tabela 3 apresenta à classificação do campeonato após 10 rodadas realizadas.

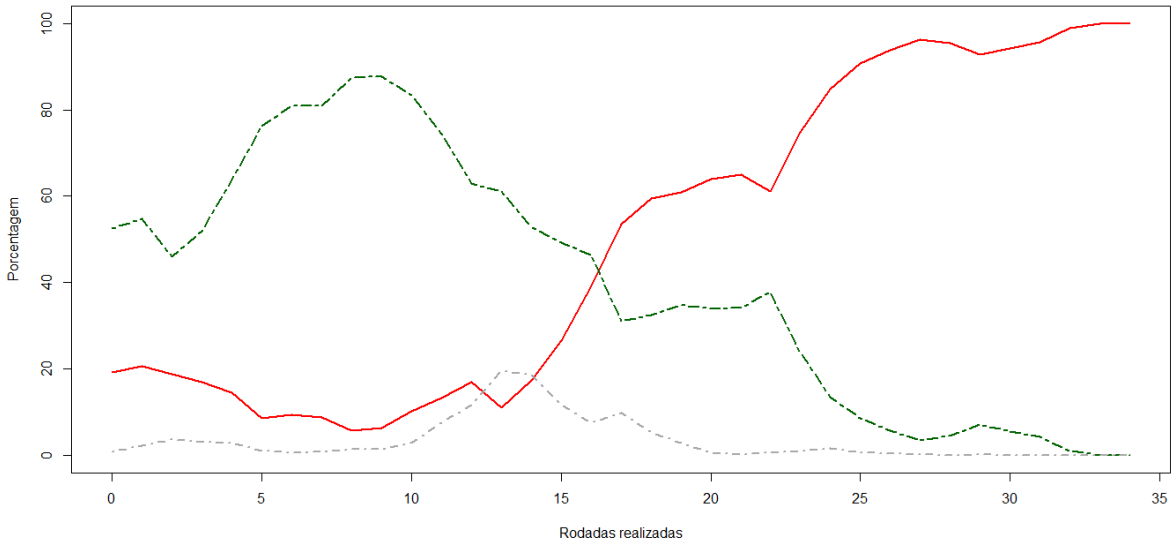
Ainda de acordo com a Tabela 3, pode-se observar que o Palmeiras possui a probabilidade de 0,8345 (ou 83,45% de chance) de ser Campeão. O Flamengo é o segundo time que possui maior probabilidade de título. Após algumas rodadas, o Palmeiras perdeu algumas partidas e, em contrapartida, o Flamengo começou a vencer as partidas disputadas. Com as mudanças dos treinadores dos clubes, os dois times permaneceram disputando o título da competição a cada rodada. A Tabela 4 apresenta o Flamengo com uma maior probabilidade de título do que o Palmeiras, devido à ascensão de vitórias do time carioca, e os tropeços realizados pelo time paulista. Por sua vez, a Tabela 5 mostra a classificação do campeonato após 34 rodadas. Esta rodada foi a decisiva para o torneio, pois o Flamengo consagrou o título da competição, restando 4 jogos para o final do campeonato.

Durante o decorrer da competição o Flamengo teve uma ascensão considerável após a vinda do seu novo treinador. A metodologia implantada pelo técnico culminou com o título de Campeão Brasileiro, assim como os destaques individuais dos jogadores. Uma forma de ilustrar esta ascensão é através da Figura 1, onde a linha vermelha refere-se ao Flamengo, a pontilhada em verde ao Palmeiras e a pontilhada e tracejada em cinza ao Santos. Vale salientar que os demais clubes da competição apresentaram um comportamento parecido ou pior que o do Santos. Deste modo, optou-se em observar apenas os três primeiros colocados da competição até a rodada 34.

De acordo com a Figura acima, pode-se observar que até a rodada 16 o Palmeiras possuía uma alta probabilidade de vencer a competição. No entanto, entre as rodadas 15 a 20 o Flamengo teve uma ascensão com sucessivas vitórias. Uma forma de justificar tal ascensão foi a vinda do novo treinador. Além disso, o Palmeiras também trocou de treinador, porém o time não conseguiu ter uma sequência de vitórias durante a troca de técnico. Outra justificativa também pode ser dada a paralisação do campeonato devido a Copa América, uma vez que o Palmeiras vinha em uma campanha excelente e após a pausa não obteve o mesmo sucesso antes da paralisação. Por outro lado, o Flamengo teve grandes resultados pós paralisação, culminando com a conquista do título.

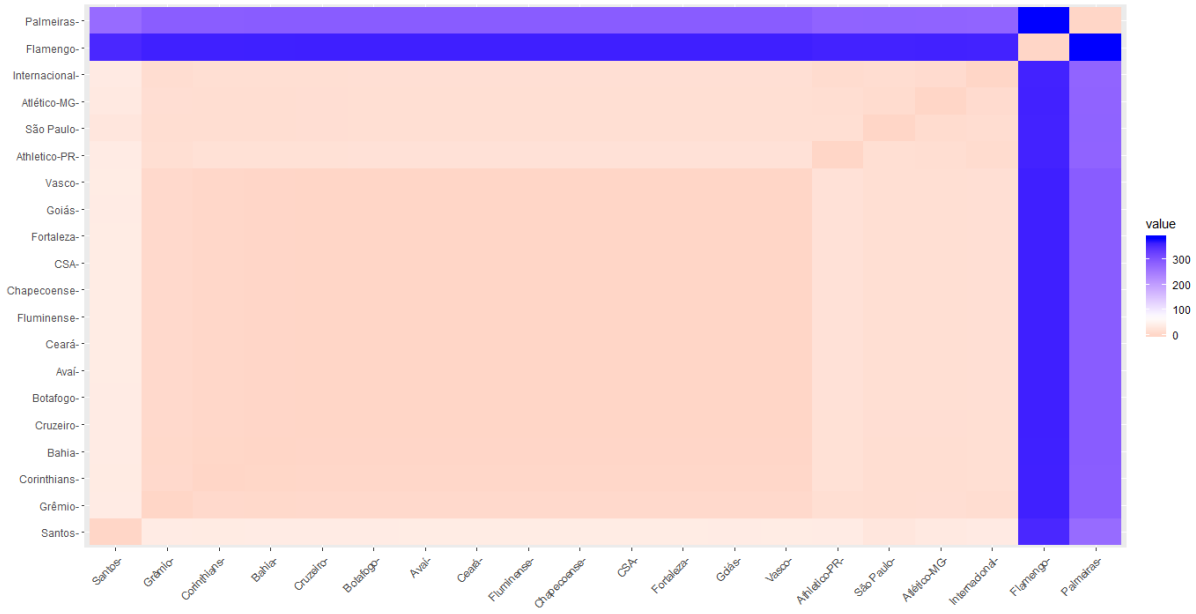
Contudo, vale salientar que Palmeiras e Flamengo foram os dois clubes que possuíam um bom plantel, além de um bom poder aquisitivo comparado aos demais clubes da competição. Neste sentido, realizou-se uma análise de agrupamento com o intuito de se obter tais conclusões. Na Análise de agrupamento, realizou-se o método hierárquico aglomerativo por ligação simples ou *Single Linkage*. Por sua vez, na matriz de similaridade utilizou-se a distância euclidiana entre os vetores das probabilidades de cada um dos vinte

Figura 1 – Gráfico referente a porcentagem dos times Flamengo, Santos e Palmeiras serem campeões ao longo das rodadas do Campeonato Brasileiro de 2019.



clubes se tornarem campeões a cada rodada. Após a obtenção da matriz de similaridade, obteve-se o Dendrograma contendo os times dispostos em grupos. A Figura 2, apresenta a matriz de similaridade entre os clubes participantes do campeonato, onde quanto mais escuro for a cor azul, mais dissimilar é o time entre os demais.

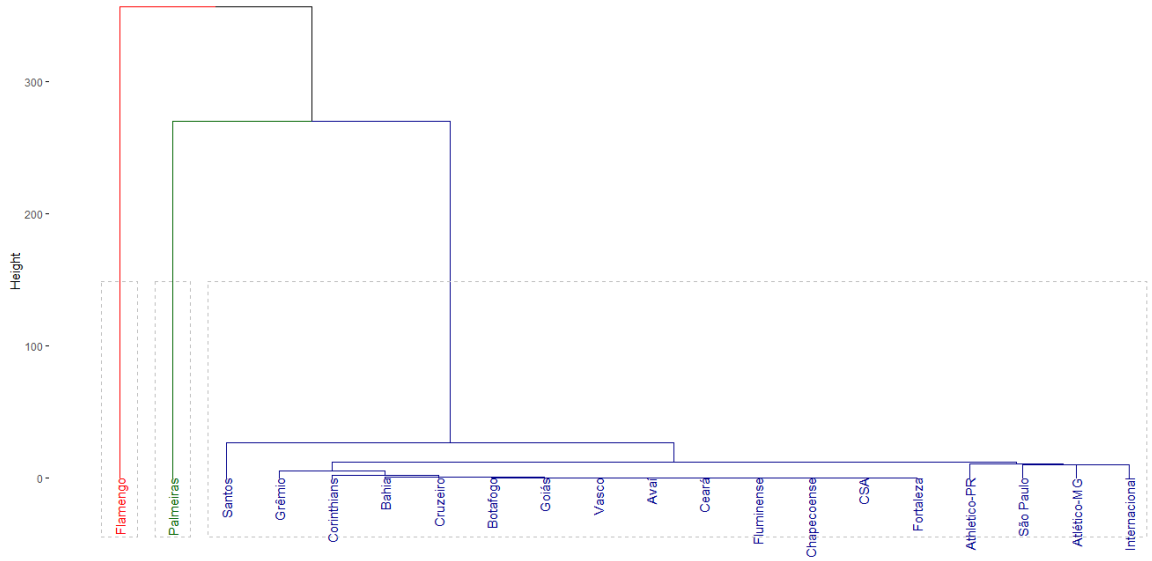
Figura 2 – Gráfico referente a Matriz de Similaridade entre os clubes do Campeonato Brasileiro de 2019.



De acordo com a Figura 2, é muito flagrante que Palmeiras e Flamengo são mais dissimilares em relação a todos os outros clubes que disputaram o campeonato. Por sua

vez, o Dendrograma, apresentado na Figura 3, apresenta os grupos nos quais os times ficaram arranjados. Neste caso obteve-se três grupos entre os 20 times da competição onde Flamengo e Palmeiras formaram um grupo exclusivo cada um, enquanto todos os 18 clubes restantes constituíram o terceiro grupo. Dessa forma, este Dendrograma representa a disparidade do Campeonato Brasileiro de 2019 quando se compara Flamengo e Palmeiras com os demais times, uma vez que ambos possuíam elencos de alto nível, bem como um poder financeiro elevado comparados aos demais clubes da competição.

Figura 3 – Dendrograma referente aos clubes participantes do Campeonato Brasileiro de 2019.



4 Conclusão

Com a popularidade da Estatística no futebol, a procura demasiada dos torcedores sobre previsões de resultados em partidas foi um dos fatores que culminaram com a implementação de modelos especializados nesta área. Existem inúmeros modelos de previsão em que pode-se discutir diversos fatores, como por exemplo o fator momento do time, isto é, se a equipe está numa crescente em um campeonato disputado, entre outros.

Por meio do Campeonato Brasileiro de 2019, aplicou-se o modelo Soma e Diferença 1, realizando simulação para as 380 partidas de forma a se alcançar a probabilidade de campeão dos times da competição. Observou-se que Palmeiras e Flamengo se destacaram durante o campeonato e assim estes disputaram a cada rodada o título do campeonato. Além disso, vale ressaltar que alguns fatores proporcionaram o rendimento destes clubes positivamente e negativamente. Por fim, por meio da análise de agrupamento ressaltou-se a grande disparidade entre as duas equipes entre as demais participantes do campeonato, haja vista que cada uma possui grandes jogadores no elenco, além de receitas consideráveis comparadas aos demais times da competição.

Tem-se a necessidade de ampliar esse estudo para diversos campeonatos, como por exemplo outros campeonatos nacionais, bem como outros regionais. Contudo, vale ressaltar que há muito que se aprender sobre esses modelos, sobretudo modelos mais robustos, como os Bayesianos, por exemplo. Dessa forma, o conhecimento adquirido é salutar para o progresso de qualquer aluno que queira adentrar nesse ramo da Estatística.

Referências

ALMUHAYFITH, F. E.; ALZAID, A. A.; OMAIR, M. A. On bivariate poisson regression models. *Journal of King Saud University-Science*, Elsevier, v. 28, n. 2, p. 178–189, 2016. Citado na página 12.

ARRUDA, M. L. d. *Poisson, Bayes, Futebol e DeFinetti*. Dissertação (Mestrado) — Universidade de São Paulo, 2000. Citado 6 vezes nas páginas 13, 14, 15, 16, 17 e 19.

BUSSAB, W. d. O.; MIAZAKI, É. S.; ANDRADE, D. d. Introdução à análise de agrupamentos. 1990. Citado na página 22.

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 12.

JOHNSON, N. L.; KEMP, A. W.; KOTZ, S. *Univariate discrete distributions*. [S.l.]: John Wiley & Sons, 2005. v. 444. Citado na página 12.

JONHSON, R. A.; WICHERN, D. W. *Applied Multivariate Statistical Analysis*. 4. ed. [S.l.]: Pearson, 1998. Citado na página 22.

KASSAMBARA, A.; MUNDT, F. *factoextra: Extract and Visualize the Results of Multivariate Data Analyses*. [S.l.], 2017. R package version 1.0.5. Disponível em: <https://CRAN.R-project.org/package=factoextra>. Citado na página 24.

LEE, A. J. Modeling scores in the premier league: is manchester united really the best? *Chance*, 1997. Citado na página 16.

MAECHLER, M. et al. *cluster: Cluster Analysis Basics and Extensions*. [S.l.], 2019. R package version 2.0.8 — For new features, see the 'Changelog' file (in the package source). Citado na página 24.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2019. Disponível em: <http://www.R-project.org/>. Citado 2 vezes nas páginas 21 e 24.

SUZUKI, A. K. *Modelagem estatística para a determinação de resultados de dados esportivos*. Dissertação (Mestrado) — Universidade Federal de São Carlos, 2007. Citado 7 vezes nas páginas 11, 12, 13, 14, 15, 17 e 18.

VICINI, L. Análise multivariada da teoria à prática. 2005. 215 f. *Monografias (especialização) Universidade Federal de Santa Maria, UFSM-Santa Maria*, 2005. Citado na página 22.

Apêndices

APÊNDICE A – Tabelas de classificação dos times no Campeonato Brasileiro de 2019

A.1 Tabelas

Este apêndice apresenta algumas tabelas obtidas a partir das simulações feitas ao longo da realização dos Campeonato Brasileiro de 2019.

Tabela 2 – Tabela referente à classificação pré-Campeonato Brasileiro de 2019.

Times	Pontos	Vitórias	Saldo Gols	Gols Pró	Probabilidade de Título
Palmeiras	0	0	0	0	0,5250
Flamengo	0	0	0	0	0,1930
Athletico-PR	0	0	0	0	0,0853
Internacional	0	0	0	0	0,0811
Grêmio	0	0	0	0	0,0533
São Paulo	0	0	0	0	0,0213
Atlético-MG	0	0	0	0	0,0186
Santos	0	0	0	0	0,0091
Bahia	0	0	0	0	0,0052
Corinthians	0	0	0	0	0,0033
Cruzeiro	0	0	0	0	0,0023
Vasco	0	0	0	0	0,0016
Avaí	0	0	0	0	0,0004
Goiás	0	0	0	0	0,0003
Botafogo	0	0	0	0	0,0001
Fluminense	0	0	0	0	0,0001
Ceará	0	0	0	0	<0,0001
Chapecoense	0	0	0	0	<0,0001
CSA	0	0	0	0	<0,0001
Fortaleza	0	0	0	0	<0,0001

Tabela 3 – Tabela referente à classificação do Campeonato Brasileiro após 10 rodadas.

Times	Pontos	Vitórias	Saldo Gols	Gols Pró	Probabilidade de título
Palmeiras	26	8	16	19	0,8345
Santos	23	7	6	13	0,0287
Flamengo	20	6	11	21	0,1016
Atlético-MG	19	6	4	16	0,0166
Corinthians	18	5	5	10	0,0020
Internacional	16	5	4	13	0,0108
Botafogo	16	5	0	8	0,0001
Goiás	15	5	-4	12	<0,0001
São Paulo	15	3	3	9	0,0017
Grêmio	14	4	0	12	0,0005
Bahia	14	4	-1	11	0,0006
Athletico-PR	13	4	2	14	0,0029
Fortaleza	13	4	-3	10	<0,0001
Ceará	11	3	1	11	<0,0001
Fluminense	9	2	-3	14	<0,0001
Cruzeiro	9	2	-7	9	<0,0001
Vasco	9	2	-7	9	<0,0001
Chapecoense	8	2	-5	11	<0,0001
CSA	6	1	-13	3	<0,0001
Avaí	4	0	-9	4	<0,0001

Tabela 4 – Tabela referente à classificação do Campeonato Brasileiro após 18 rodadas.

Times	Pontos	Vitórias	Saldo Gols	Gols Pró	Probabilidade de Título
Flamengo	39	12	23	41	0,5653
Santos	37	11	12	30	0,0515
Palmeiras	36	10	15	29	0,3592
Corinthians	32	8	10	21	0,0033
São Paulo	31	8	8	20	0,0071
Internacional	30	9	6	22	0,0075
Bahia	30	8	6	21	0,0014
Atlético-MG	27	8	3	23	0,0015
Athletico-PR	26	8	7	25	0,0031
Botafogo	26	8	-1	18	<0,0001
Grêmio	25	6	2	24	0,0001
Ceará	21	6	1	21	<0,0001
Fortaleza	21	6	-4	21	<0,0001
Goiás	21	6	-12	17	<0,0001
Vasco	20	5	-9	16	<0,0001
Cruzeiro	18	4	-11	16	<0,0001
Fluminense	15	4	-9	20	<0,0001
CSA	15	3	-17	7	<0,0001
Chapecoense	14	3	-14	16	<0,0001
Avaí	10	1	-16	9	<0,0001

Tabela 5 – Tabela referente à classificação do Campeonato Brasileiro após 34 rodadas.

Times	Pontos	Vitórias	Saldo Gols	Gols Pró	Probabilidade de título
Flamengo	81	25	43	73	1
Santos	68	20	23	53	0
Palmeiras	68	19	26	53	0
Grêmio	59	17	23	57	0
Athletico-PR	56	16	16	47	0
São Paulo	54	14	9	34	0
Internacional	51	14	5	39	0
Corinthians	50	12	6	36	0
Goiás	46	13	-15	39	0
Bahia	44	11	1	39	0
Vasco	44	11	-6	36	0
Fortaleza	43	12	-2	44	0
Atlético-MG	41	11	-6	39	0
Botafogo	39	12	-12	29	0
Fluminense	38	10	-10	34	0
Ceará	37	10	-1	33	0
Cruzeiro	36	7	-13	27	0
CSA	29	7	-30	21	0
Chapecoense	28	6	-21	27	0
Avaí	18	3	-36	16	0