



Universidade de Brasília
Departamento de Estatística

Estimação do número de gols esperados em jogos do campeonato brasileiro
de 2024 via modelo Poisson

Daniel Paranaguá Fontenelle Fraga

Relatório apresentado para o Departamento de Estatística da Universidade de Brasília como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Brasília
2025

Daniel Paranaguá Fontenelle Fraga

**Estimação do número de gols esperados em jogos do campeonato brasileiro
de 2024 via modelo Poisson**

Orientador(a): Prof(a). Leandro Tavares Correia

Relatório apresentado para o Departamento de Estatística da Universidade de Brasília como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

**Brasília
2025**

Agradecimentos

Gostaria de expressar minha sincera gratidão ao meu orientador, Leandro Tavares, pelo valioso auxílio, orientação e paciência ao longo da realização deste trabalho. Sua dedicação e comprometimento comigo foram fundamentais para o desenvolvimento deste projeto. Agradeço também à minha família, que sempre esteve ao meu lado durante toda a trajetória deste curso, oferecendo apoio e incentivo nos momentos mais difíceis. Por fim, deixo um agradecimento especial aos meus amigos Arthur Rodrigues, Davi Esmeraldo, Gabriela Lôbo, Lucas Menezes e Matheus Martinez, que enfrentaram comigo os desafios dessa jornada acadêmica desde os primeiros semestres, com muita parceria envolvida e noites sem dormir compartilhadas. Este trabalho e todo meu desenvolvimento é fruto de todas as vivências que compartilhamos ao longo desses anos.

Resumo

O presente trabalho tem como objetivo principal estimar o número de gols esperados em partidas do Campeonato Brasileiro de Futebol de 2024 por meio de modelos estatísticos de contagem. Para isso, foram utilizados dados das rodadas já realizadas na competição, com ênfase nos gols marcados por mandantes e visitantes. A metodologia baseia-se em dois tipos principais de modelos: o modelo de Poisson Bivariado, com diferentes estruturas para a covariância entre os gols das equipes, e o modelo de regressão de Poisson. A análise permitiu identificar os modelos mais adequados com base em critérios como AIC, BIC e aderência à distribuição de Poisson. Além disso, foram feitas previsões de confrontos específicos, avaliando-se a performance preditiva dos modelos em termos de acurácia dos resultados previstos, com destaque para a influência do mando de campo na estimativa dos gols esperados e nas probabilidades de vitória, empate ou derrota. Os resultados obtidos mostram que os modelos propostos são eficazes na modelagem e previsão de gols, fornecendo insights relevantes para aplicações em análises esportivas.

Palavras-chaves: Campeonato Brasileiro, Modelo Poisson, Gols Esperados, Modelos Bivariados, Previsão de Resultados.

Lista de Tabelas

1	Estatísticas descritivas dos gols marcados por localização	23
2	Forças ofensivas (α) e defensivas (β) estimadas via modelo M12 (Athletico Paranaense como referência)	26
3	Resumo do desempenho dos times no Brasileirão 2024.	28
4	Comparação entre os modelos bivariados ajustados segundo AIC e BIC. . .	33
5	Previsões dos confrontos mais desequilibrados em termos de probabilidade.	35
6	Gols observados e esperados na 1ª rodada do Campeonato Brasileiro de 2024	37

Lista de Figuras

1	Distribuição dos gols observados e simulados por Poisson, por localização . . .	24
2	Gráfico das forças ofensivas (α) e defensivas (β) estimadas para os times da Série A (modelo M12).	27
3	Mapa de calor dos resultados dos confrontos entre mandantes e visitantes no Brasileirão 2024.	30
4	Comparação dos modelos segundo os critérios AIC e BIC.	34
5	Probabilidades previstas de resultado na 1ª rodada do Brasileirão 2024. . .	38

Sumário

1 Introdução	8
2 Referencial Teórico	10
2.1 Modelo de Regressão Linear	10
2.2 Modelos Lineares Generalizados	11
2.3 Modelo de Poisson	13
2.4 Modelo de Poisson Bivariado	15
3 Metodologia	17
3.1 Conjunto de dados	17
3.1.1 Variáveis de interesse	17
3.1.2 Obtenção dos dados	18
3.1.3 Reprodutibilidade e organização do código	18
3.2 Ajuste e Seleção de Modelos Bivariados	19
3.3 Procedimento de Previsão	20
3.4 Avaliação das Previsões	22
3.4.1 Probabilidade do resultado real	22
3.4.2 Previsão do desfecho da partida	22
4 Resultados	23
4.1 Análise Descritiva dos Gols	23
4.2 Forças Ofensivas e Defensivas	25
4.3 Desempenho Geral dos Times	28
4.4 Mapa de Resultados dos Confrontos	29
4.5 Ajuste dos Modelos de Poisson	31
4.6 Modelagem por Regressão de Poisson	32
4.7 Ajuste e Comparação dos Modelos de Poisson Bivariado	33
4.8 Previsão de Confrontos com o Modelo de Poisson	35
4.9 Avaliação da Performance Preditiva	36

5 Conclusão	39
Referências.	41
Apêndice	42

1 Introdução

O Futebol é um esporte que teve sua origem na Inglaterra, no ano de 1863. Apesar de naquela época, o esporte não ter o formato, as regras ou sequer o nome que conhecemos atualmente, o futebol começou a se expandir de maneira gradual pelo país, até se tornar a paixão global que conhecemos hoje e, a pergunta "Para qual time você torce?" se tornar tão comum quanto perguntar a alguém "Qual a sua profissão?" ou "Quais são os seus hobbies?", por exemplo. Atualmente, o futebol conta com mais de 3 bilhões de torcedores no mundo inteiro, sendo o esporte mais assistido e jogado de todos, transcendendo fronteiras sociais e culturais e sendo amplamente reconhecido pelo seu impacto em diversos âmbitos.

Conhecido por seu enorme apelo popular, o futebol tem sido foco de interesse científico em diversas áreas, como estatística, econometria e ciência dos dados. Os Modelos estatísticos têm sido ferramentas importantíssimas para entender e quantificar diversos eventos relacionados ao esporte, como previsão de resultados, análise de desempenho, acontecimentos durante a partida e comportamento tático. Estudos baseados em modelos lineares, séries temporais e distribuições de probabilidade, como a Poisson, têm explorado diferentes abordagens para quantificar a previsibilidade de resultados esportivos.

Casas de apostas esportivas, como a Betano, a VaideBet e a Betfair, por exemplo, desempenham um papel importante no aumento do interesse da população que é fã do esporte por previsões quantitativas no mundo futebolístico. Uma das métricas mais populares é a previsão do número de gols esperados em uma partida, que possui aplicações tanto no cenário acadêmico quanto no mercado financeiro e esportivo. Trabalhos recentes como os de Cordeiro e Demétrio (2008) destacam a relevância de Modelos Lineares Generalizados para modelar dados de contagem, enquanto Paula (2024) aborda extensões bivariadas que permitem analisar relações entre variáveis correlacionadas, como os gols marcados por times adversários.

Ademais, modelos clássicos como o de Maher (1982), por exemplo, introduziram a utilização da distribuição de Poisson para modelar o número de gols em partidas de futebol, destacando sua aplicação em previsões esportivas. Por outro lado, em estudos um pouco mais recentes, como Dixon e Coles (1997), foi aprimorada essa abordagem ao incluir a análise temporal e ajustar as taxas de ocorrência com base em fatores que dependem do contexto, como o mando de campo, o histórico entre os clubes e a fase atual de cada time. Essas contribuições estabeleceram uma base para o uso de modelos estatísticos ao se analisar resultados esportivos.

Ademais, modelos clássicos como o de Maher (1982), por exemplo, introduziram a utilização da distribuição de Poisson para modelar o número de gols em partidas de futebol, destacando sua aplicação em previsões esportivas. Por outro lado, em estudos um pouco mais recentes, como Dixon e Coles (1997), foi aprimorada essa abordagem ao incluir a análise temporal e ajustar as taxas de ocorrência com base em fatores que dependem do contexto, como o mando de campo, o histórico entre os clubes e a fase atual de cada time. Além disso, Karlis e Ntzoufras (2005) propuseram a modelagem conjunta dos gols do mandante e do visitante por meio da distribuição de Poisson bivariada, o que permite incorporar a dependência entre os gols das duas equipes e capturar aspectos não tratados pelos modelos univariados, como a correlação entre os desempenhos ofensivo e defensivo em uma mesma partida. Essas contribuições estabeleceram uma base para o uso de modelos estatísticos ao se analisar resultados esportivos

Este trabalho tem como objetivo modelar e prever o número de gols em partidas do Campeonato Brasileiro Série A de 2024, por meio de modelos de Poisson e Poisson bivariado. Será analisada a capacidade do modelo de prever com precisão a quantidade de gols marcados e sofridos pelos times que participaram deste campeonato, contribuindo para um maior entendimento da dinâmica do esporte e fornecendo ferramentas aplicáveis para diversas áreas, como análises esportivas e apostas. Para isso, foram formulados os seguintes objetivos específicos: Construir um banco de dados contendo os resultados das partidas da Série A do Campeonato Brasileiro de 2024, aplicar diferentes variações do modelo de Poisson Bivariado, com foco na modelagem dos gols marcados pelas equipes mandantes e visitantes, comparar os modelos em termos de desempenho e avaliar a capacidade preditiva do modelo selecionado, por meio da comparação entre os gols esperados e os resultados observados.

O trabalho está organizado da seguinte forma: na seção 2, apresenta-se o referencial teórico sobre o modelo de Poisson, sua aplicação em partidas de futebol, assim como a abordagem mais recente deste modelo. A seção 3 descreve os dados utilizados, o processo de coleta e os métodos estatísticos empregados na análise. Em seguida, a seção 4 apresenta os resultados obtidos, incluindo uma análise descritiva, ajuste e comparação dos modelos, bem como as previsões realizadas. Por fim, na seção 5, são discutidas as principais conclusões do estudo e sugestões para trabalhos futuros.

2 Referencial Teórico

Nesta seção, apresenta-se o referencial teórico que fundamenta a análise estatística dos resultados esportivos com foco na modelagem do número de gols em partidas de futebol. Inicialmente, são abordados os conceitos de regressão linear, modelo base para muitas análises estatísticas e ponto de partida para o entendimento dos modelos mais complexos. Em seguida, discute-se a estrutura dos modelos lineares generalizados (MLG), que estendem a regressão linear para lidar com variáveis resposta que seguem distribuições distintas da normal, como a Poisson.

A partir disso, detalha-se o modelo de Poisson e sua aplicação para as variáveis de contagem, como o número de gols em partidas. Ademais, é explorado o modelo de Poisson bivariado, que permite considerar a possível dependência entre os gols marcados pelas equipes mandante e visitante.

Esses tópicos fornecem a base teórica necessária para a construção e avaliação dos modelos utilizados na análise dos dados do Campeonato Brasileiro de 2024.

2.1 Modelo de Regressão Linear

O modelo de regressão linear é usado para modelar a relação entre uma variável resposta contínua Y e uma ou mais variáveis explicativas $X_1, X_2, \dots, X_p$. O modelo assume que a relação entre as variáveis é linear e que os erros são normalmente distribuídos com média igual a zero e variância constante.

Sua fórmula geral é dada por:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \varepsilon,$$

em que $\beta_0, \beta_1, \dots, \beta_p$ são os coeficientes do modelo e ε representa o erro aleatório. O termo ε representa parte da variabilidade de Y que não é explicada pelas variáveis explicativas, ou seja, engloba os efeitos de fatores não observados ou não incluídos no modelo.

Para que o modelo de regressão linear produza estimativas confiáveis por meio do método dos mínimos quadrados ordinários (MQO), é necessário que as seguintes suposições clássicas sejam atendidas:

- **Linearidade:** a relação entre a variável resposta e as variáveis explicativas deve ser linear nos parâmetros;
- **Independência dos erros:** os erros devem ser estatisticamente independentes entre si;
- **Homoscedasticidade:** os erros devem ter variância constante para todos os valores das variáveis explicativas;
- **Normalidade dos erros:** os erros devem seguir uma distribuição normal com média zero;
- **Ausência de multicolinearidade:** as variáveis explicativas não devem ser linearmente dependentes entre si.

O entendimento adequado dessas suposições é essencial, pois sua violação pode comprometer a validade inferencial do modelo e a precisão das estimativas dos parâmetros (MONTGOMERY; PECK; VINING, 2021).

2.2 Modelos Lineares Generalizados

Os Modelos Lineares Generalizados (MLG) são uma extensão dos modelos de regressão linear que permitem lidar com diferentes distribuições da variável resposta Y , como a binomial, a Poisson, a gama, entre outras (PAULA, 2024).

Os MLGs são compostos por três elementos fundamentais:

- **Família Exponencial:** A variável resposta Y deve seguir uma distribuição pertencente à família exponencial, cuja função densidade de probabilidade (ou massa de probabilidade, no caso discreto) pode ser representada na forma:

$$f(y; \theta, \phi) = \exp \left(\frac{y\theta - b(\theta)}{\phi} + c(y, \phi) \right),$$

em que:

- θ é o parâmetro canônico da distribuição;
- ϕ é o parâmetro de dispersão;
- $b(\theta)$ é uma função que garante a normalização da distribuição;
- $c(y, \phi)$ é uma função que depende apenas de y e ϕ , também necessária para a normalização.

Esta fórmula inclui diversas distribuições comuns, como normal, binomial e Poisson.

- **Estrutura do MLG:** Os MLGs mantêm um preditor linear, assim como na regressão linear, mas incorporam uma função de ligação que conecta a média da variável resposta $\mu = \mathbb{E}[Y]$ ao preditor linear. A estrutura geral é:

$$\eta = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p, \quad g(\mu) = \eta,$$

onde η é o preditor linear e $g(\cdot)$ é a função de ligação. A escolha da função de ligação depende da distribuição da variável resposta. Por exemplo, usa-se a função logarítmica para a distribuição de Poisson.

- **Processo de Estimação:** Os parâmetros β são estimados por meio do *método da máxima verossimilhança*. Esse método busca encontrar os valores de β que maximizam a probabilidade de observar os dados disponíveis. A função de verossimilhança para uma amostra de tamanho n é definida como:

$$L(\beta; \mathbf{y}) = \prod_{i=1}^n f(y_i; \beta),$$

e a log-verossimilhança correspondente é:

$$\ell(\beta; \mathbf{y}) = \sum_{i=1}^n \log f(y_i; \beta).$$

Os estimadores de máxima verossimilhança são obtidos ao resolver o sistema de equações:

$$\frac{\partial \ell(\beta)}{\partial \beta} = 0,$$

que, sob condições regulares, garante que os estimadores sejam consistentes, assintoticamente normais e eficientes.

- **Pressupostos:** Para a correta aplicação dos MLGs, devem ser satisfeitos os seguintes pressupostos:
 - A variável resposta Y segue uma distribuição da família exponencial;
 - Existe uma relação linear entre o preditor η e os preditores $X_1, \dots, X_p$;
 - A função de ligação $g(\cdot)$ é monotônica e diferenciável, garantindo a conexão entre μ e η ;
 - As observações Y_i , com $i \in \{1, 2, \dots, n\}$, são independentes entre si;
 - A variância de Y pode ser expressa como uma função da média, isto é, $\text{Var}(Y) = \phi V(\mu)$, onde $V(\mu)$ é chamada função de variância.

2.3 Modelo de Poisson

A distribuição de Poisson é usada para modelar contagens, como o número de gols em uma partida, por exemplo. A sua função de probabilidade é dada por:

- Função de Probabilidade:

$$P(Y = y) = \exp\{y \log(\mu) - \mu - \log(y!)\}, \quad \text{com } \mu > 0 \text{ e } y = 0, 1, 2, \dots$$

- Em que: $\log(\mu) = \theta$, $b(\theta) = e^\theta$, $\phi = 1$ e $c(y; \phi) = -\log(y!)$.
- Esperança e Variância:

- Esperança: $\mathbb{E}[Y] = \mu$;
- Variância: $\text{Var}(Y) = \mu$.

Esses momentos podem ser obtidos a partir da função geradora de momentos da distribuição de Poisson ou, de forma mais geral, a partir da forma da função da família exponencial, conforme discutido anteriormente. Para a distribuição de Poisson, que pertence à família exponencial, sabemos que:

$$b'(\theta) = \mu \quad \text{e} \quad b''(\theta) = \text{Var}(Y)$$

o que confirma que $\mathbb{E}[Y] = \mu$ e $\text{Var}(Y) = \mu$.

O modelo de Poisson é amplamente utilizado para modelar dados de contagem, sendo adequado em situações nas quais a média e a variância da variável resposta são iguais, o que constitui uma restrição fundamental desse modelo (PAULA, 2024). A função de verossimilhança é escrita em termos de λ , que representa o parâmetro da distribuição de Poisson. Como, nesse modelo, a esperança e a variância são iguais a μ , temos que $\lambda = \mu$, e a utilização de λ neste contexto segue a notação clássica da distribuição de Poisson. Essa função de verossimilhança, dado um conjunto de observações $y_1, y_2, \dots, y_n$, é dada por:

$$L(\lambda) = \prod_{i=1}^n \frac{\lambda^{y_i} e^{-\lambda}}{y_i!},$$

onde:

- y_i é a observação de contagem para o i -ésimo caso,
- λ é o parâmetro de Poisson.

A função de ligação mais comum para o modelo de Poisson é a **função logarítmica**, que relaciona a média da variável resposta $\mu_i = \mathbb{E}[Y_i]$ com o preditor linear η_i , sendo dada por:

$$\log(\mu_i) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p.$$

Essa função garante que a média da variável resposta μ seja sempre positiva, o que é essencial no contexto de variáveis de contagem, como o número de gols ou o número de eventos em um intervalo fixo. Outras funções de ligação que podem ser usadas no modelo de Poisson incluem:

- **Função Identidade:** Outra função de ligação possível é a identidade, onde a média μ_i é igual ao preditor linear:

$$g(\mu_i) = \mu_i.$$

Embora seja uma função de ligação válida, não é comumente utilizada para modelos de Poisson, pois permite que a média μ_i assumam valores negativos, não sendo adequado para variáveis de contagem.

- **Função Inversa:** A função inversa é uma outra possibilidade, mas também é menos comum no modelo de Poisson:

$$g(\mu_i) = \frac{1}{\mu_i}.$$

- **Função Proporcional:** Em alguns casos, a função de ligação proporcional pode ser utilizada:

$$g(\mu_i) = \mu_i^a \quad (a \text{ é uma constante, e define a forma da função de ligação}).$$

- Se $a = 1$, tem-se a **função identidade**, ou seja, $g(\mu_i) = \mu_i$;
- Se $a = -1$, tem-se a **função inversa**, isto é, $g(\mu_i) = \mu_i^{-1}$.

Essa função é útil quando se deseja modelar a média da variável de contagem como uma razão proporcional.

2.4 Modelo de Poisson Bivariado

O modelo de Poisson bivariado permite modelar duas variáveis de contagem correlacionadas. Esse modelo é particularmente útil para capturar a dependência entre os processos de contagem (SILVA, 2014). A probabilidade conjunta das variáveis X e Y é dada por:

$$P(X = x, Y = y) = e^{-(\lambda_1 + \lambda_2 + \lambda_3)} \frac{\lambda_1^x \lambda_2^y}{x! y!} \sum_{k=0}^{\min(x, y)} \binom{x}{k} \binom{y}{k} k! \left(\frac{\lambda_3}{\lambda_1 \lambda_2} \right)^k.$$

Os parâmetros do modelo possuem as seguintes interpretações, no contexto da modelagem do número de gols esperados:

- λ_1 : Taxa média de gols esperados do time mandante de forma independente.
- λ_2 : Taxa média de gols esperados do time visitante de forma independente.
- λ_3 : Captura a dependência entre os processos de contagem dos dois times.

Além disso, as esperanças e a covariância das variáveis são expressas por:

$$\mathbb{E}[X] = \lambda_1 + \lambda_3, \quad \mathbb{E}[Y] = \lambda_2 + \lambda_3, \quad \text{Cov}(X, Y) = \lambda_3.$$

A estimação dos parâmetros do modelo é feita pelo método de máxima verossimilhança. A função de verossimilhança é:

$$L(\lambda_1, \lambda_2, \lambda_3; k_1, k_2) = \prod_{i=1}^n \sum_{n=0}^{\min(k_{1i}, k_{2i})} \frac{\lambda_1^{k_{1i}-n} \lambda_2^{k_{2i}-n} \lambda_3^n e^{-(\lambda_1 + \lambda_2 + \lambda_3)}}{(k_{1i} - n)!(k_{2i} - n)!n!}.$$

onde k_{1i} e k_{2i} representam, respectivamente, os valores observados das variáveis de contagem para a i -ésima observação.

O somatório interno, que varia de $n = 0$ até $\min(k_{1i}, k_{2i})$, reflete a convolução entre as componentes compartilhadas da contagem (associadas a λ_3) e as partes específicas de cada variável (relacionadas a λ_1 e λ_2). Esse termo é característico do modelo de Poisson Bivariado e garante que a soma seja feita sobre todas as possíveis combinações de decomposição dos totais observados.

Esse modelo é amplamente utilizado para prever resultados de jogos, pois, considera tanto a taxa média de gols de cada time, quanto a interação entre eles, permitindo uma modelagem mais realista da distribuição dos gols em partidas de futebol (PAULA, 2024).

Em algumas aplicações, observa-se uma frequência de empates maior do que a prevista pelo modelo de Poisson bivariado tradicional. Para lidar com esse fenômeno, introduz-se a *inflação na diagonal*, que consiste em aumentar a probabilidade associada aos eventos em que os dois times marcam o mesmo número de gols, ou seja, $X = Y$. Essa modificação melhora o ajuste do modelo em contextos onde empates ocorrem com frequência superior à esperada (KARLIS; NTZOUFRAS, 2005).

3 Metodologia

Diante da complexidade e limitações encontradas na obtenção das variáveis de desempenho adicionais (como número de chutes, defesas, posse de bola, escanteios...), optou-se por direcionar a análise exclusivamente para a modelagem dos gols marcados, tanto pelo time mandante quanto pelo time visitante.

Essa abordagem se fundamenta em utilizar a Distribuição de Poisson Bivariada para prever resultados das partidas de futebol a partir do histórico de gols marcados. Como cada partida pode ser representada por um par de contagens inteiras (gols do time mandante e gols do time visitante), os modelos baseados em contagem, como o modelo de Poisson e suas extensões, são adequados para capturar a estrutura dos dados. Além disso, ao se restringir apenas a variável de gols, a análise torna-se mais robusta e replicável, uma vez que os dados de gols marcados pelos times estão amplamente disponíveis e possuem um menor viés de coleta em comparação a outras estatísticas de desempenho, que podem variar dependendo da fonte utilizada.

A modelagem baseia-se, portanto, em estruturas que utilizam os gols como resposta e incluem apenas os efeitos fixos dos times mandantes e visitantes na predição do número esperado de gols em uma partida. Diversos modelos foram testados, incluindo o modelo de Poisson Bivariado padrão, variações com inflação de empates, modelos com diferentes estruturas para o parâmetro de dependência λ_3 , e modelos inflados na diagonal.

3.1 Conjunto de dados

O banco de dados utilizado neste estudo foi construído exclusivamente a partir do pacote `worldfootballR`, que permite acessar dados esportivos diretamente de fontes confiáveis como *FBref* e *Transfermarkt*. A coleta considerou todas as 380 partidas disputadas no Campeonato Brasileiro Série A de 2024.

A principal motivação para essa escolha foi a facilidade de acesso aos dados organizados e a possibilidade de obter, de forma padronizada, os placares reais das partidas e demais variáveis de interesse.

3.1.1 Variáveis de interesse

O conjunto de dados foi construído com as seguintes variáveis:

1. **JogoID**: Identificador único da partida (seguindo a ordem do campeonato).
2. **Turno**: Indica se a partida ocorreu no primeiro ou no segundo turno.
3. **Mandante**: Nome da equipe mandante.
4. **Visitante**: Nome da equipe visitante.
5. **GolsMando**: Quantidade de gols marcados pelo time mandante.
6. **GolsFora**: Quantidade de gols marcados pelo time visitante.
7. **xG_Mandante**: Número de gols esperados (expected goals) do time mandante.
8. **xG_Visitante**: Número de gols esperados (expected goals) do time visitante.

3.1.2 Obtenção dos dados

Essa fonte de dados utilizada foi escolhida por sua ampla cobertura das competições nacionais e internacionais, pela atualização frequente e pela estrutura consistente dos dados. Além disso, a padronização das variáveis e a facilidade de acesso via API do `worldfootballR` tornam essa fonte confiável e conveniente para análises estatísticas. A extração considerou todas as partidas do campeonato brasileiro de 2024, garantindo informações completas e atualizadas sobre os gols marcados e os valores esperados de gols de cada time.

3.1.3 Reprodutibilidade e organização do código

O código foi estruturado de forma modular, com scripts separados para as seguintes etapas:

- Coleta e limpeza dos dados;
- Ajuste dos modelos;
- Geração de previsões;
- Cálculo de métricas de avaliação;
- Geração de gráficos e tabelas.

A organização buscou seguir as práticas comuns de programação e análise estatística, permitindo uma fácil reprodutibilidade do trabalho e adaptação para outras competições ou temporadas futuras.

3.2 Ajuste e Seleção de Modelos Bivariados

Com o conjunto de dados estruturado e a definição da variável resposta focada exclusivamente nos gols marcados por mandantes e visitantes, o passo seguinte foi analisar o ajuste de diferentes versões do modelo de Poisson Bivariada. A ideia central foi encontrar uma estrutura de modelagem que melhor capturasse a dependência entre os dois processos de contagem, ao mesmo tempo em que fornecesse uma boa predição.

A modelagem foi feita utilizando a função `lm.bp()` do pacote `Bivpois`, que permite ajustar modelos bivariados da família de Poisson em diferentes configurações para os parâmetros $\lambda_1, \lambda_2, \lambda_3$. No presente trabalho, esses parâmetros são estimados sem o uso de covariáveis, ou seja, assumem valores constantes para todas as observações. Foram testadas diferentes estruturas para este parâmetro, inclusive modelos com intercepto comum, interceptos diferenciados por time e modelos inflacionados na diagonal (KARLIS; NTZOUFRAS, 2005).

A seguir, descreve-se brevemente os principais modelos ajustados:

- **Modelo M1 (Básico):** modelo bivariado padrão com intercepto comum, sem considerar efeitos específicos dos times.
- **Modelo M2 (Efeitos fixos):** inclusão de efeitos fixos para mandante e visitante, mantendo o intercepto comum.
- **Modelo M3 (Efeitos fixos com $\lambda_3 \sim 1$):** considera os efeitos fixos de mandante e visitante com interceptos distintos e λ_3 constante.
- **Modelo M4 (Efeitos fixos com $\lambda_3 \sim 1$, intercepto comum):** similar ao M3, mas com intercepto comum entre os efeitos de mandante e visitante.
- **Modelo M5 (Efeitos fixos com $\lambda_3 = 0$):** modelo que assume independência entre os gols do mandante e visitante (sem termo de correlação).
- **Modelo M6 (Inflado na diagonal):** modelo com inflação de empates, inspirado na estrutura proposta por Karlis e Ntzoufras (2005), visando capturar uma maior ocorrência de empates nas partidas.

- **Modelo M7 (Com intercepto livre para λ_3):** versão mais flexível, permitindo que o parâmetro λ_3 varie livremente com o intercepto.
- **Modelo M8 (Com intercepto comum para λ_3):** assume um único intercepto para λ_3 , exigindo uma maior parcimônia.
- **Modelo M9 (Sem intercepto e com λ_3 variável):** modelo que exclui o intercepto comum e permite a variação livre de λ_3 .
- **Modelo M10 (Efeitos aditivos para times e $\lambda_3 \sim 1$):** inclui o somatório de efeitos fixos por time mandante e visitante com λ_3 constante.
- **Modelo M11 (Com estrutura completa e λ_3 livre):** modelo mais completo estimado, com interceptos diferenciados para mandante, visitante e estrutura livre para λ_3 , sem restrições.
- **Modelo M12 (Completo Penalizado):** Modelo mais completo, com penalização, mando de campo, λ_3 e inflação na diagonal.

A seleção do modelo mais adequado foi feita com base em critérios de verossimilhança, como o valor da log-verossimilhança maximizada, AIC (Akaike Information Criterion), e o comportamento gráfico das distribuições ajustadas em comparação com os dados observados. Os modelos com estruturas mais flexíveis para o parâmetro λ_3 tenderam a capturar melhor a correlação entre os gols, especialmente nos casos em que jogos equilibrados influenciam simultaneamente os desempenhos ofensivos de ambos os times.

3.3 Procedimento de Previsão

Após a estimação dos modelos bivariados, o próximo passo consiste em utilizar os parâmetros estimados para realizar previsões das partidas do Campeonato Brasileiro. O objetivo é estimar, para cada confronto, o número de gols esperados que o time mandante e a equipe visitante deverão marcar, além de calcular as probabilidades associadas aos diferentes desfechos possíveis do jogo (como vitória do mandante, empate ou vitória do visitante), a partir das probabilidades conjuntas de gols geradas pela distribuição de Poisson Bivariada.

A previsão foi realizada utilizando os valores estimados de λ_1 , λ_2 e λ_3 para cada partida da base de dados, conforme o modelo escolhido. A função massa de probabilidade da distribuição de Poisson Bivariada foi usada para obter as probabilidades conjuntas de todos os pares possíveis de gols (de 0 a 5) para as equipes mandantes e visitantes. Esse limite superior de 5 gols foi adotado com base na análise empírica dos dados, em que a maioria esmagadora dos placares observados no Campeonato Brasileiro apresentam até 5 gols por equipe. Assim, essa truncagem visa balancear a representação realista dos dados com a redução da complexidade computacional do modelo, visto que eventos com 6 ou mais gols são raros e têm pouca influência nas probabilidades conjuntas estimadas.

O procedimento seguiu os seguintes passos principais:

1. **Extração dos parâmetros estimados:** a partir do modelo final ajustado, foram obtidos os valores estimados dos parâmetros de ataque e defesa dos times, bem como do termo de dependência λ_3 .
2. **Cálculo dos gols esperados:** para cada partida i , foram calculados:

$$\hat{\mu}_{\text{mandante},i} = \lambda_1^{(i)}, \quad \hat{\mu}_{\text{visitante},i} = \lambda_2^{(i)}, \quad \text{com dependência } \lambda_3^{(i)}.$$

3. **Geração das distribuições conjuntas:** utilizando a fórmula da Poisson Bivariada, foi construída a matriz de probabilidades

$$P(X = x, Y = y) \quad \text{para } x, y = 0, 1, \dots, 5,$$

representando as chances de o jogo terminar com x gols do mandante e y gols do visitante.

4. **Cálculo das probabilidades de resultado:** com a matriz gerada, foram somadas as probabilidades associadas a cada tipo de resultado:

- $P(\text{Vitória do Mandante}) = \sum_{x>y} P(x, y)$
- $P(\text{Empate}) = \sum_{x=y} P(x, y)$
- $P(\text{Vitória do Visitante}) = \sum_{x<y} P(x, y)$

5. **Construção da tabela de previsão:** os resultados foram organizados em uma tabela com os seguintes campos:

- Times (mandante e visitante);
- Gols esperados para cada equipe;
- Probabilidades associadas a cada desfecho possível da partida (vitória do mandante, empate ou vitória do visitante).

3.4 Avaliação das Previsões

Para os jogos analisados, realizou-se uma avaliação da performance preditiva do modelo selecionado, com base na comparação entre os resultados previstos e os resultados efetivamente observados.

A avaliação foi feita por meio de diferentes abordagens, conforme recomendado por Karlis e Ntzoufras (2005) e também presente no trabalho de Almeida (2024), que exploram o uso do modelo de Poisson Bivariado não apenas para estimativa de parâmetros, mas também como ferramenta de previsão confiável no contexto esportivo.

3.4.1 Probabilidade do resultado real

Para cada jogo, foi calculada a probabilidade do resultado real observado (placar exato) segundo a distribuição de Poisson Bivariada parametrizada com os valores estimados. A média dessas probabilidades ao longo das partidas fornece uma medida geral da verossimilhança preditiva do modelo. Valores mais altos indicam que o modelo, mesmo quando não acerta exatamente o placar, atribui probabilidades significativas para os resultados que de fato ocorrem.

3.4.2 Previsão do desfecho da partida

Após o cálculo da probabilidade do resultado real, foi avaliado se o modelo foi capaz de prever corretamente o desfecho da partida, ou seja, se o time mandante venceu, houve empate ou se o time visitante venceu. O modelo atribui probabilidades a cada um desses três possíveis desfechos, e foi considerado um acerto quando o desfecho com maior probabilidade coincidiu com o resultado real da partida.

4 Resultados

4.1 Análise Descritiva dos Gols

Antes de fazer a aplicação dos modelos estatísticos, foi realizada uma análise descritiva dos gols marcados nas partidas do Campeonato Brasileiro de 2024. O objetivo foi entender o comportamento das variáveis de interesse e verificar se a distribuição dos gols pode ser considerada compatível com o pressuposto do modelo de Poisson.

Tabela 1: Estatísticas descritivas dos gols marcados por localização

Local	Média	Mediana	Moda
Mandante	1,41	1,4	2
Visitante	1,03	1,02	1

A Tabela 1 apresenta as estatísticas descritivas dos gols marcados por times mandantes e visitantes. Observa-se que os times mandantes marcaram, em média, mais gols do que as equipes visitantes, o que pode indicar a presença de uma vantagem em jogar em casa. Além disso, nota-se que, para ambos os grupos, as três medidas são próximas, sugerindo uma distribuição assimétrica moderada, mas ainda compatível com o pressuposto de média igual à variância, conforme o modelo de Poisson. Para os times mandantes, a média foi de 1,41 gols, a mediana de 1,4 e a moda foi de 2, enquanto para os visitantes os valores foram 1,03, 1,02 e 1, respectivamente, sugerindo que, de acordo com as 3 medidas, o time mandante tende a marcar mais gols que o time visitante. Também foi realizada uma simulação com base em distribuições de Poisson ajustadas às médias empíricas, a fim de comparar o comportamento dos dados observados com os dados simulados.

Para visualizar o comportamento das distribuições empíricas, foram gerados gráficos comparando as frequências de gols observadas com distribuições simuladas segundo uma Poisson com mesma média. O Gráfico 1 ilustra essa comparação, separadamente, para as equipes mandantes e visitantes.

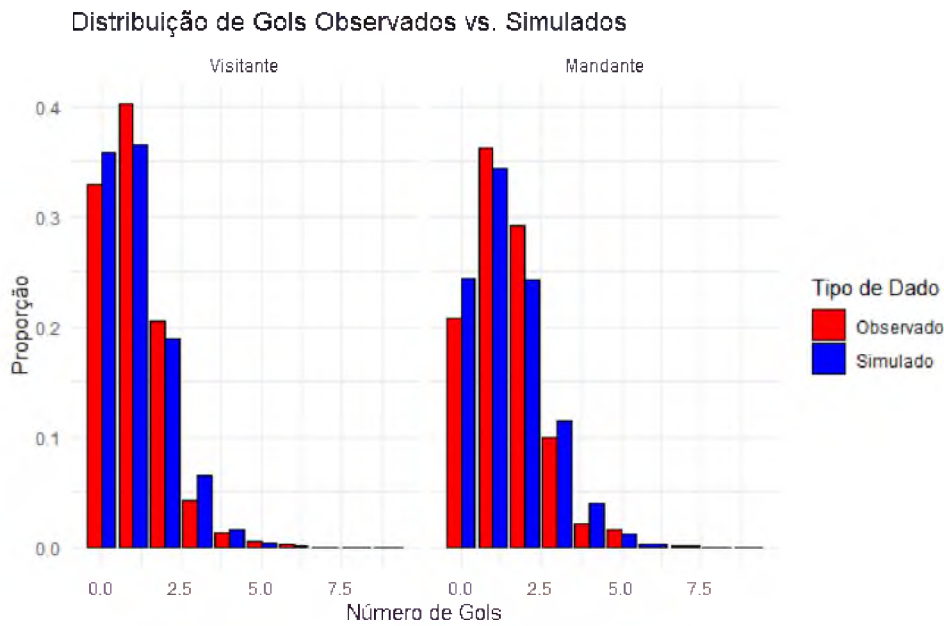


Figura 1: Distribuição dos gols observados e simulados por Poisson, por localização

A Figura 1 apresenta a distribuição dos gols observados e simulados para os times mandantes e visitantes, considerando as médias empíricas obtidas no banco de dados como um parâmetro para a geração da distribuição de Poisson. Visualmente, nota-se que as distribuições observadas e simuladas são bastante parecidas, especialmente em relação aos gols mais frequentes (0, 1 e 2 gols), tanto para mandantes quanto para visitantes.

Para os times visitantes, observa-se que a maior parte das partidas resultou em 0 ou 1 gol, o que também é refletido na distribuição simulada. Já para os mandantes, é visível uma leve tendência a marcar mais gols, com dois picos mais evidentes nas distribuições: um próximo de 1 gol e outro de 2 gols, indicando uma característica multimodal no comportamento do número de gols marcados. Essa assimetria entre mandantes e visitantes é coerente com os valores médios encontrados (1,41 para mandantes e 1,03 para visitantes), e reforça a presença de uma vantagem ao jogar em casa.

Nas extremidades da distribuição (número de gols mais altos), os dados simulados apresentaram uma ligeira superestimação em relação aos observados. Ainda assim, a proximidade entre as distribuições comparadas sugere que o modelo de Poisson, mesmo em sua forma mais simples, consegue representar razoavelmente bem o comportamento empírico das contagens de gols.

Por fim, para verificar a aderência dos dados à distribuição de Poisson, foi aplicado o teste de aderência do tipo qui-quadrado. Esse teste foi conduzido separadamente para os gols de equipes mandantes e visitantes. Os resultados são apresentados a seguir:

- Gols dos mandantes: $p\text{-valor} = 0,084$
- Gols dos visitantes: $p\text{-valor} = 0,069$

Os valores de p -valor indicam que não há evidências suficientes para rejeitar a hipótese de que os gols seguem uma distribuição de Poisson, tanto para os times mandantes, quanto para os times visitantes. Essa constatação é coerente com o esperado, conforme sugerido em Karlis e Ntzoufras (2005), na aplicação prática realizada por Almeida (2024) e nas simulações apresentadas em R-bloggers (2019) e R-bloggers (2020). Com base nesses resultados apresentados, é razoável prosseguir com a modelagem baseada na distribuição de Poisson.

4.2 Forças Ofensivas e Defensivas

Após a análise descritiva dos gols, foi construída uma tabela com a estimativa das forças ofensivas e defensivas dos times participantes do Campeonato Brasileiro de 2024. Essa etapa busca compreender melhor o desempenho dos clubes com base na modelagem dos gols esperados, levando em conta a variação da performance ao jogar em casa ou fora. Para isso, utilizou-se o Modelo M12, um modelo de Poisson Bivariado Penalizado com efeitos de ataque (α) e defesa (β) específicos para cada time. Esse modelo permite separar o desempenho ofensivo e defensivo das equipes e leva em consideração a inflação na diagonal da matriz de probabilidades, sendo adequado para lidar com a dependência entre os gols marcados por mandantes e visitantes.

A Tabela 2 apresenta os valores estimados das forças ofensivas (α) e defensivas (β) para 19 dos 20 clubes da Série A. O Athletico Paranaense foi utilizado como time de referência, tendo seus parâmetros fixados em zero para garantir a identificação do modelo, o que é uma prática comum em modelos com efeitos fixos. Isso não significa que o time tenha desempenho nulo ou neutro dentro do campeonato, mas apenas que os valores dos demais times são interpretados em relação a ele.

Valores positivos indicam, para a força ofensiva, uma tendência a marcar mais gols; e, para a força defensiva, uma tendência a sofrer mais gols. Já valores negativos indicam menor capacidade do ataque em marcar gols ou uma maior solidez defensiva.

Tabela 2: Forças ofensivas (α) e defensivas (β) estimadas via modelo M12 (Athletico Paranaense como referência)

Time	Força Ofensiva (α)	Força Defensiva (β)
Atlético Goianiense	-0,220	0,309
Atlético Mineiro	-0,169	-0,171
Bahia	-0,074	-0,207
Botafogo	0,441	-0,370
Corinthians	0,006	-0,300
Criciúma	-0,285	-0,066
Cruzeiro	0,112	-0,067
Cuiabá	-0,051	0,319
Flamengo	0,067	-0,419
Fluminense	0,173	0,201
Fortaleza	0,151	-0,274
Grêmio	-0,088	-0,100
Internacional	0,231	-0,271
Juventude	-0,259	-0,198
Palmeiras	0,310	-0,392
RB Bragantino	-0,047	-0,098
São Paulo	0,053	-0,279
Vasco da Gama	-0,201	-0,084
Vitória	-0,129	-0,125

A Figura 2 apresenta o gráfico de barras com as forças ofensivas e defensivas estimadas. As barras azuis representam o parâmetro ofensivo (α) e as barras vermelhas, o parâmetro defensivo (β). O gráfico permite visualizar, de forma direta, quais times possuem maior capacidade de ataque e quais sofrem mais gols, segundo o modelo.

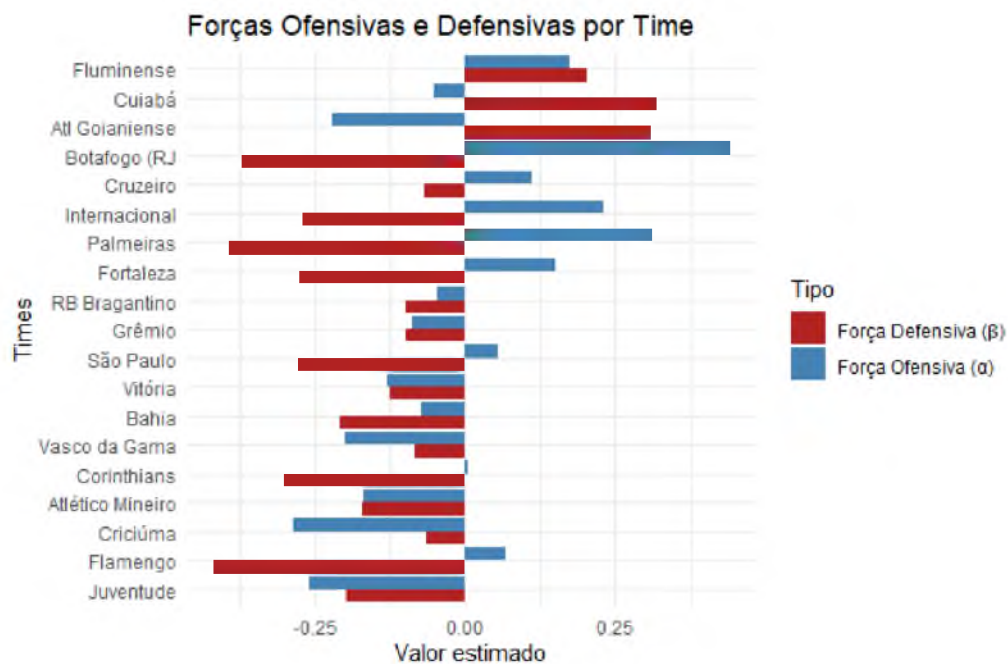


Figura 2: Gráfico das forças ofensivas (α) e defensivas (β) estimadas para os times da Série A (modelo M12).

Analisando a Tabela 2 e a Figura 2 nota-se que, times com maiores valores de α , como Botafogo, Palmeiras e Internacional, possuem uma maior força ofensiva estimada. Além disso, do ponto de vista defensivo, os times com menores valores de β , como Flamengo, Palmeiras e São Paulo, mostraram maior solidez defensiva, sofrendo menos gols segundo o modelo ajustado. Ademais, a combinação dos dois parâmetros permite análises mais completas. Por exemplo, o RB Bragantino apresenta valores próximos de zero para ambos os parâmetros, o que pode indicar um desempenho bastante equilibrado.

Por fim, é importante ressaltar que os parâmetros estimados pelo modelo não representam médias simples de gols, mas sim valores ajustados a partir da estrutura estatística do modelo de Poisson Bivariado. Assim, algumas discrepâncias entre os valores estimados e a classificação final dos clubes podem ocorrer, como, por exemplo, a equipe do Flamengo, que teve o melhor ataque do campeonato de 2024, mas isso não se refletiu nos valores ajustados. o que reforça a utilidade dos modelos estatísticos como ferramenta complementar à análise bruta dos dados.

Essas estimativas servem como base para a próxima etapa da análise, em que serão feitas previsões detalhadas dos confrontos, incluindo a probabilidade de vitória, empate ou derrota em cada partida.

4.3 Desempenho Geral dos Times

A Tabela 3 resume o desempenho dos times no Campeonato Brasileiro de 2024, considerando os pontos conquistados, gols marcados, gols sofridos e o saldo de gols. Essa análise fornece uma visão geral dos resultados obtidos pelas equipes ao longo da competição e permite comparações diretas entre seus desempenhos ofensivo e defensivo.

Tabela 3: Resumo do desempenho dos times no Brasileirão 2024.

Time	Pontos	Gols Marcados	Gols Sofridos	Saldo
Botafogo	79	59	29	+30
Palmeiras	73	60	33	+27
Flamengo	70	61	42	+19
Fortaleza	68	53	39	+14
Internacional	65	53	36	+17
São Paulo	59	53	43	+10
Corinthians	56	54	45	+9
Bahia	53	49	49	0
Cruzeiro	52	43	41	+2
Vasco da Gama	50	43	56	-13
Vitória	47	45	52	-7
Atlético Mineiro	47	47	54	-7
Fluminense	46	33	39	-6
Juventude	45	48	59	-11
Grêmio	45	44	50	-6
RB Bragantino	44	44	48	-4
Ath Paranaense	42	40	46	-6
Criciúma	38	42	61	-19
Atl Goianiense	30	29	58	-29
Cuiabá	30	29	49	-20

Observa-se que o Botafogo finalizou a competição com a melhor campanha em termos de pontuação, com 79 pontos, seguido por Palmeiras, com 73 pontos, e pelo Flamengo, com 70. Além disso, o Botafogo também foi o time com o maior saldo de gols no campeonato, com um saldo igual a +30, evidenciando uma boa consistência ofensiva e defensiva. Já o Flamengo, por sua vez, foi o time com maior número de gols marcados, com 61 gols, reforçando seu grande poder ofensivo.

Na parte inferior da tabela, os times com maior fragilidade defensiva foram o Atlético Goianiense, com 58 gols sofridos, o Criciúma, com 61 e o Juventude, com 59. O Atlético Goianiense, além disso, teve o pior saldo de gols da competição, com um saldo de -29, consequência de uma combinação entre um ataque pouco eficiente, com apenas 29 gols marcados, e uma defesa bastante vulnerável, o que ocasionou a sua baixa classificação.

Apesar da tabela refletir os dados brutos da competição, ela complementa a análise estatística baseada em modelos. Um exemplo claro disso é o contraste entre os valores observados e os estimados pelo modelo M12, pois, mesmo que o Flamengo seja o time com maior ataque, isso não se refletiu na estimação obtida pelo modelo. Isso ocorre, pois, o modelo ajusta os parâmetros de ataque e defesa conforme o contexto específico de cada confronto, considerando o adversário, o mando de campo e a estrutura bivariada dos gols, não apenas as médias globais.

Esse contraste ressalta a importância de interpretar os resultados dos modelos estatísticos em conjunto com os dados observados. A abordagem combinada oferece uma visão mais rica e detalhada sobre o desempenho de cada equipe ao longo da temporada, evitando conclusões precipitadas e permitindo uma compreensão mais precisa dos fatores que influenciam os resultados.

4.4 Mapa de Resultados dos Confrontos

A Figura 3 apresenta um mapa de calor contendo os resultados dos confrontos entre os times ao longo do Campeonato Brasileiro de 2024. As linhas representam os times mandantes e as colunas, os times visitantes. Cada célula mostra o placar do jogo com o mandante à esquerda e o visitante à direita.

A coloração das células facilita a visualização dos resultados:

- Verde: vitória do mandante;
- Vermelho: vitória do visitante;
- Cinza: empate.

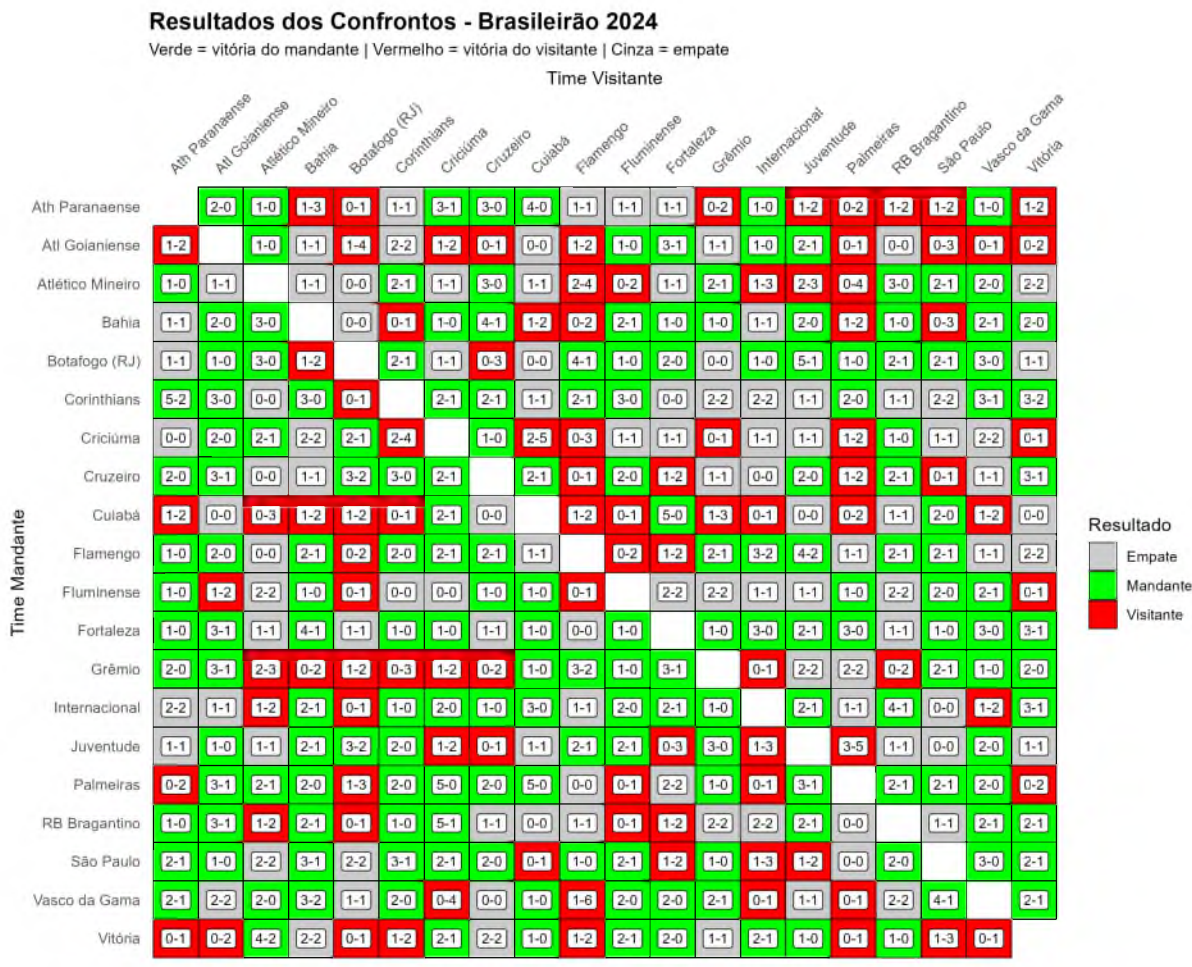


Figura 3: Mapa de calor dos resultados dos confrontos entre mandantes e visitantes no Brasileirão 2024.

Observa-se na Figura 3 que o Botafogo venceu a maior parte dos jogos em casa, enquanto o Cuiabá e o Atlético Goianiense sofreram diversas derrotas, mesmo como mandantes. Por outro lado, equipes como Flamengo e Palmeiras conseguiram vitórias importantes fora de casa, destacando sua consistência mesmo em condições mais adversas.

Esse tipo de visualização permite identificar padrões importantes, como a dominância em casa, desempenho equilibrado fora, ou dificuldade específica contra determinados adversários. Além disso, oferece uma visão complementar às análises anteriores baseadas em tabelas e modelos estatísticos.

A matriz também revela confrontos mais equilibrados, onde ocorreram empates, tanto no primeiro, quanto no segundo turno, além de resultados mais elásticos, que reforçam o bom desempenho ofensivo de alguns clubes.

De maneira geral, observa-se uma tendência de vantagem para os times mandantes, que se reflete na predominância de células verdes na diagonal inferior da matriz. Entretanto, alguns clubes destoaram dessa regra, conquistando mais pontos como visitantes, o que pode indicar uma preparação tática mais eficaz fora de casa ou uma maior capacidade de adaptação. Já times com poucas vitórias em ambas as condições evidenciam uma fragilidade técnica ou instabilidade ao longo do campeonato inteiro.

4.5 Ajuste dos Modelos de Poisson

Para dar continuidade à modelagem estatística dos gols, foi ajustado o modelo M12, com o objetivo de estimar o número esperado de gols marcados nas partidas do Campeonato Brasileiro de 2024. Esse modelo, implementado por meio da função `lm.dibp()` da biblioteca `dibp` no R, é uma extensão do Modelo de Poisson Bivariado, permitindo estimar separadamente os gols dos times mandantes e visitantes com base em três efeitos principais: o fator casa, a força ofensiva do time e a força defensiva do adversário.

Além de captar essas variações, o modelo se mostrou adequado por possibilitar a realização de previsões individuais para os gols de cada equipe, a partir dos parâmetros estimados.

O modelo foi ajustado com a seguinte estrutura:

```
ex4.m12 <- lm.dibp(HomeGoals ~ 1, AwayGoals ~ 1, l1l2 = form1,  
                  data = brasileirao_matches,  
                  distribution = "poisson",  
                  zeroL3 = TRUE)
```

Nesse caso, o argumento `zeroL3 = TRUE` define que o parâmetro de correlação entre os gols marcados pelo time mandante e o time visitante (terceiro parâmetro do modelo bivariado) é fixado em zero. Isso significa que esse modelo assume independência condicional entre os gols das duas equipes, o que simplifica a estrutura do modelo e o torna mais estável para fins de previsão.

A especificação `l1l2 = form1` indica que os efeitos de ataque e defesa dos times variam conforme o mando de campo. Dessa forma, cada time possui parâmetros distintos de força ofensiva e defensiva, que interagem com a condição de ser mandante ou visitante dentro do confronto.

4.6 Modelagem por Regressão de Poisson

Com base nos parâmetros estimados a partir do modelo M12, que forneceu as forças ofensivas (α), defensivas (β) e o fator casa (γ) —, foi possível construir um modelo de regressão de Poisson simples para representar o número esperado de gols em cada confronto.

Nessa abordagem, a quantidade de gols esperados por cada time em uma partida é modelada por uma estrutura aditiva, típica de regressão de Poisson, considerando três componentes principais:

- **Fator casa (γ):** Representa a vantagem média de jogar como mandante;
- **Força ofensiva do time (α):** Indica a capacidade média de marcar gols;
- **Força defensiva do adversário (β):** Reflete a tendência do oponente em sofrer gols.

A partir desses parâmetros, o número esperado de gols ($\mathbb{E}(Gols)$) para cada equipe em uma partida foi estimado com a fórmula:

$$\mathbb{E}(Gols) = \exp(\gamma + \alpha_{\text{time}} - \beta_{\text{adversário}})$$

Essas estimativas foram utilizadas como entrada em um modelo de regressão de Poisson univariado, implementado com a função `glm()` no R, para calcular as distribuições de gols esperados por time.

Posteriormente, as probabilidades de vitória do mandante, empate e vitória do visitante foram obtidas com a função `prever_prob_resultado()`, que calcula a convolução das distribuições marginais de gols de cada equipe, assumindo independência condicional.

As análises e previsões apresentadas nas subseções seguintes demonstram a aplicação prática desse modelo preditivo, oferecendo uma ferramenta quantitativa para avaliação de cenários prováveis em confrontos reais.

Essa abordagem, baseada na modelagem estatística dos gols, permite não apenas compreender os desempenhos passados, mas também realizar projeções fundamentadas em evidências históricas e no contexto específico de cada partida.

4.7 Ajuste e Comparação dos Modelos de Poisson Bivariado

Nesta etapa, foram ajustados doze modelos de Poisson Bivariado (denominados M1 a M12), com diferentes estruturas envolvendo efeitos de ataque, defesa, mando de campo e dependência entre os gols das equipes. O objetivo foi avaliar qual configuração oferece o melhor desempenho na explicação dos dados observados do Campeonato Brasileiro de 2024.

A Tabela 4 a seguir, apresenta os valores dos critérios de informação AIC (Akaike Information Criterion) e BIC (Bayesian Information Criterion), utilizados para comparar a qualidade dos modelos penalizando pela complexidade. Valores menores desses critérios indicam melhor ajuste com menor risco de sobreajuste.

Tabela 4: Comparação entre os modelos bivariados ajustados segundo AIC e BIC.

Modelo	AIC	BIC
M1	2103,27	2288,60
M2	2105,27	2295,24
M3	2137,92	2415,92
M4	2134,75	2412,75
M5	2151,64	2517,68
M6	2107,27	2301,87
M7	2109,27	2308,50
M8	2108,06	2307,29
M9	2105,32	2309,19
M10	2107,32	2315,82
M11	2109,27	2308,50
M12	2107,27	2301,87

Observa-se que o modelo com o menor valor de AIC foi o **M1**, seguido de perto pelos modelos **M2**, **M6**, **M12** e **M10**. Em relação ao BIC, os melhores desempenhos também foram observados para os modelos **M1**, **M2** e **M6**, indicando que essas especificações oferecem um bom ajuste com menor penalização pela complexidade.

Além dos critérios numéricos, um fator importante na escolha do modelo é a **inflação na diagonal**, ou seja, a capacidade de capturar a frequência elevada de empates em partidas de futebol. O modelo **M12**, apesar de não apresentar o menor AIC ou BIC absolutos, possui uma estrutura que contempla essa característica ao permitir a modelagem conjunta dos gols, separadamente, para mandantes e visitantes, com efeitos diferenciados.

Considerando todos os critérios, será mantido o modelo **M12** nas análises seguintes, tanto pela sua performance satisfatória quanto pela estrutura flexível que ele oferece.

A Figura 4 ilustra os valores de AIC e BIC para todos os modelos, facilitando a visualização das diferenças de desempenho entre os modelos avaliados.

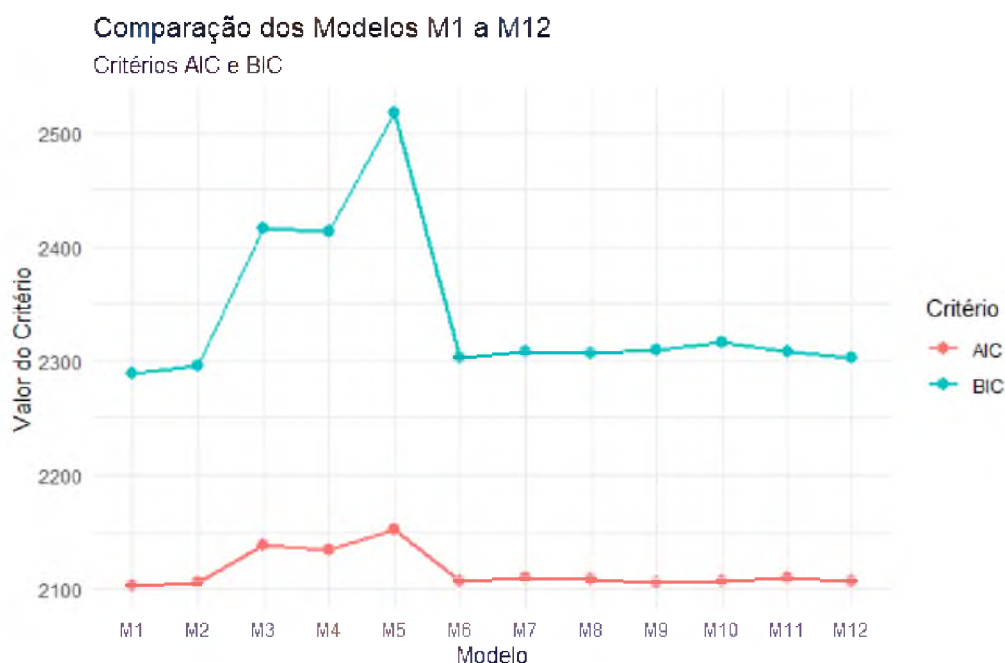


Figura 4: Comparação dos modelos segundo os critérios AIC e BIC.

A Figura 4 complementa os dados da Tabela 4, ao ilustrar visualmente os valores dos critérios AIC e BIC para cada um dos modelos ajustados. Observa-se uma diferença considerável no BIC dos modelos M3, M4 e M5, que apresentam valores significativamente mais altos do que os demais, indicando que esses modelos são penalizados por sua maior complexidade em comparação com o ganho de ajuste.

Por outro lado, os demais modelos destacam-se por apresentarem valores baixos tanto para o AIC, quanto para o BIC, o que os torna boas opções em termos de parcimônia e desempenho preditivo.

A análise gráfica também reforça a escolha do modelo M12, que aparece entre os melhores colocados nos dois critérios e, ao mesmo tempo, oferece uma formulação que equilibra qualidade de ajuste, estrutura interpretável e aplicabilidade nas previsões.

4.8 Previsão de Confrontos com o Modelo de Poisson

A partir das estimativas do modelo de regressão de Poisson ajustado via `glm()`, foram geradas previsões para diferentes confrontos do Campeonato Brasileiro de 2024. O modelo considera os efeitos de mando de campo, força ofensiva e força defensiva das equipes e, foi feito por meio da função `prever_prob_resultado()`, que retorna o número esperados de gols para o time mandante e visitante, bem como as probabilidades associadas a cada resultado possível: vitória do mandante, empate ou vitória do visitante.

De forma mais precisa, os gols esperados são obtidos com base na predição do modelo `glm()`, e as probabilidades dos desfechos são calculadas a partir da distribuição de Poisson assumida para cada time. Portanto, os resultados apresentados nesta seção não são derivados do modelo M12, mas sim do modelo de Poisson simples ajustado com `glm()`.

A Tabela 5 apresenta os dez confrontos mais desequilibrados, ordenados pela diferença absoluta entre a maior e a menor probabilidade de resultado. Na Tabela, “Gols E. M.” e “Gols E. V.” expressam a quantidade de gols esperados por mandante e visitante em casa confronto, respectivamente, enquanto que “Vitória M.” e “Vitória V.” expressam a probabilidade de vitória das equipes mandante e visitante, respectivamente.

Tabela 5: Previsões dos confrontos mais desequilibrados em termos de probabilidade.

Mandante	Visitante	Gols E.M	Gols E.V	Vitória M	Empate	Vitória V
Palmeiras	Atlético-GO	2,17	0,62	0,83	0,11	0,06
Flamengo	Atlético-GO	2,11	0,70	0,79	0,13	0,08
Botafogo	Cuiabá	2,00	0,72	0,77	0,14	0,09
Palmeiras	Criciúma	1,94	0,81	0,76	0,15	0,09
Internacional	Atlético-GO	1,83	0,75	0,75	0,17	0,08
Atlético-GO	Palmeiras	0,62	2,17	0,11	0,06	0,83
Cuiabá	Botafogo	0,69	2,00	0,13	0,15	0,72
Criciúma	Flamengo	0,81	2,11	0,15	0,17	0,68
Vitória	Palmeiras	0,85	1,92	0,17	0,21	0,62
Bahia	Flamengo	0,91	1,84	0,20	0,19	0,61

A análise dos confrontos mais desequilibrados revela padrões interessantes em termos de favoritismo. Destaca-se o confronto entre Palmeiras e Atlético-GO, em que o time mandante apresenta um número esperado de 2,17 gols, contra apenas 0,62 do visitante, o que se reflete em uma probabilidade de 83% de vitória para o Palmeiras. Esse mesmo confronto, quando invertido (Atlético-GO como mandante), apresenta a mesma expectativa de gols, porém com o favoritismo se deslocando para o visitante, evidenciando a robustez da força ofensiva e defensiva do Palmeiras em ambos os cenários.

Outro exemplo marcante é o confronto entre Flamengo e Atlético-GO. A expectativa de gols de 2,11 para o Flamengo contra 0,70 para o Atlético-GO, o que se traduz em uma probabilidade de vitória de 79% para o rubro-negro, destacando o desequilíbrio técnico estimado pelo modelo.

Essas previsões refletem as forças ofensivas e defensivas estimadas de cada time, além do fator casa, conforme discutido nas seções anteriores. A inclusão dos gols esperados na análise permite uma interpretação mais rica dos resultados, evidenciando que os desfechos com maior probabilidade estão diretamente relacionados às diferenças entre os gols médios projetados para cada equipe.

A abordagem fornece, assim, uma ferramenta útil para simulações e análises de favoritismo em diferentes rodadas da competição, permitindo antecipar o comportamento provável dos resultados com base em dados históricos e parâmetros estatísticos consolidados.

4.9 Avaliação da Performance Preditiva

A partir das estimativas do modelo de regressão de Poisson ajustado via `glm()`, o objetivo nessa etapa foi comparar os gols esperados com os gols efetivamente observados em uma rodada específica do Campeonato Brasileiro de 2024. Para isso, foram selecionadas as partidas da primeira rodada da competição.

A Tabela 6 apresenta os placares reais observados, os gols esperados para mandante e visitante segundo o modelo e, se houve ou não acerto de previsão. Consideramos que houve acerto quando o time com o maior número de gols esperados coincide com o vencedor da partida. Em casos de empate, considerou-se acerto se os valores de gols esperados também foram próximos (empate predito). Assim como na tabela anterior, “Gols E. M.” e “Gols E. V.” expressam a quantidade de gols esperados por mandante e visitante, respectivamente, enquanto que “Resultado E.” seria o resultado esperado da partida.

Tabela 6: Gols observados e esperados na 1ª rodada do Campeonato Brasileiro de 2024

Mandante vs Visitante	Placar	Gols E. M.	Gols E. V.	Resultado E.	Acerto?
Vitória vs Palmeiras	0x1	0,99	1,47	Visitante	Sim
Cruzeiro vs Botafogo	3x2	0,82	1,13	Empate	Não
Atlético-GO vs Flamengo	1x2	0,82	1,66	Visitante	Sim
Corinthians vs Atlético-MG	0x0	1,91	1,03	Mandante	Não
Vasco vs Grêmio	2x1	1,42	1,18	Empate	Não
Athletico-PR vs Cuiabá	4x0	1,26	0,64	Mandante	Sim
Fluminense vs Bragantino	2x2	1,02	0,81	Empate	Sim
São Paulo vs Fortaleza	1x2	1,36	1,09	Empate	Não
Internacional vs Bahia	2x1	1,69	0,85	Mandante	Sim
Criciúma vs Juventude	1x1	1,65	1,42	Empate	Sim

A partir da Tabela 6, observa-se que o modelo acertou o resultado esperado em 6 dos 10 confrontos. Isso indica que, mesmo com um modelo simples baseado apenas em gols e fatores fixos, é possível capturar uma parte bastante relevante da dinâmica dos jogos.

Além disso, nota-se que casos como o jogo entre Corinthians e Atlético-MG mostram limitações do modelo: apesar do time mandante ter uma expectativa de gols significativamente maior (1.91 contra 1.03), o empate entre as equipes prevaleceu. Em contrapartida, jogos como Internacional x Bahia e Atlético-PR x Cuiabá demonstram excelente desempenho preditivo do modelo.

De forma geral, os resultados reforçam a utilidade do modelo de Poisson como uma ferramenta viável para estimativas e simulações no futebol. Ainda que não preveja com exatidão todos os placares, o modelo apresenta coerência com a lógica competitiva e fornece estimativas próximas da realidade observada.

Como reforço visual, a Figura 5 apresenta as probabilidades previstas para cada um dos desfechos (vitória do mandante, empate e vitória do visitante) nos confrontos da primeira rodada.

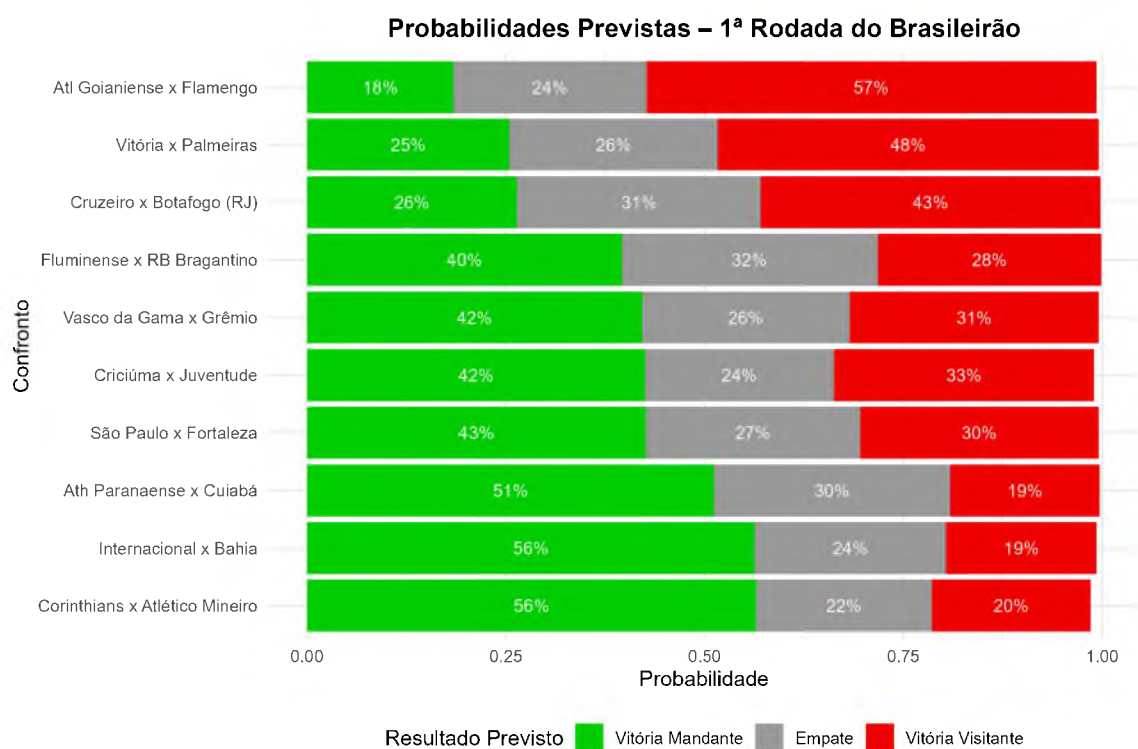


Figura 5: Probabilidades previstas de resultado na 1ª rodada do Brasileirão 2024.

A Figura 5 apresenta, de forma visual, as probabilidades previstas de vitória do mandante, empate e vitória do visitante para os confrontos da 1ª rodada do Campeonato Brasileiro de 2024. Observa-se que, para a maioria dos jogos, as maiores probabilidades estão associadas à vitória do time mandante, com destaque para Corinthians, Internacional e Athletico Paranaense. No entanto, há confrontos em que a vitória do visitante possui maior probabilidade, como no caso de Atlético-GO contra Flamengo e Vitória contra Palmeiras. O gráfico reforça visualmente o padrão de favoritismo indicado pelo modelo ajustado, considerando o desempenho ofensivo e defensivo estimado de cada equipe e o fator casa.

5 Conclusão

Este trabalho teve como objetivo principal estimar o número de gols esperados em partidas do Campeonato Brasileiro de 2024, por meio da aplicação de modelos estatísticos baseados na distribuição de Poisson. A abordagem adotada priorizou a simplicidade e interpretabilidade dos modelos, permitindo avaliar se os dados empíricos de gols podem ser considerados compatíveis com os pressupostos teóricos dessa distribuição.

Inicialmente, a análise descritiva mostrou que os times mandantes tendem a marcar mais gols do que os visitantes, sugerindo a presença de uma vantagem associada ao mando de campo. A partir disso, foram ajustados modelos de regressão de Poisson via `glm()` para estimar os gols esperados, e modelos bivariados (M1 a M12) para capturar relações mais complexas entre os gols marcados por mandantes e visitantes, considerando diferentes estruturas de dependência e possíveis inflacionamentos na diagonal.

Os resultados mostraram que, mesmo com sua estrutura mais simples, o modelo de regressão de Poisson ajustado via `glm()` foi capaz de gerar previsões coerentes com os dados observados. A análise preditiva indicou que, em uma parte considerável dos confrontos avaliados, os gols esperados estiveram próximos dos gols efetivamente realizados, com acertos significativos na previsão do placar em confrontos reais.

Ademais, os modelos bivariados foram avaliados com base em critérios como AIC, BIC e presença de inflação na diagonal, sendo o modelo M12 o que apresentou o melhor desempenho comparativo. Embora o modelo M12 não tenha sido utilizado para a previsão direta dos gols esperados, ele foi fundamental para a análise estrutural e comparação entre diferentes hipóteses sobre o comportamento conjunto dos gols de mandantes e visitantes.

Por fim, destaca-se que o uso da distribuição de Poisson mostrou-se adequado para modelar contagens de gols, especialmente pelo fato dos dados terem sido tratados e analisados com cuidado e os parâmetros estimados foram interpretados de maneira condicional ao adversário enfrentado e à condição de mando. A metodologia aqui aplicada pode ser estendida para outras temporadas, bem como aprimorada com a inclusão de variáveis explicativas adicionais (como posse de bola, finalizações, defesas e etc.), contribuindo para análises mais robustas e precisas no campo das estatísticas aplicadas ao futebol.

Os resultados obtidos estão em consonância com as conclusões apresentadas por Karlis e Ntzoufras (2005), que destacam a utilidade do modelo de Poisson Bivariado na modelagem de resultados em esportes, especialmente quando há possível correlação entre as contagens. Além disso, a estratégia metodológica aqui empregada foi fortemente

inspirada nos procedimentos descritos por Almeida (2024), que utilizou como base a construção feita em R-bloggers (2019) e R-bloggers (2020), o que reforça a consistência e aplicabilidade dos modelos utilizados neste trabalho.

Referências

- ALMEIDA, L. M. d. *Previsão de resultados de partidas da série A do Campeonato Brasileiro de Futebol : aplicações do modelo de Poisson*. 2024. <<https://lume.ufrgs.br/handle/10183/277000>>. Trabalho de Conclusão de Curso (Graduação em Matemática) – Universidade Federal do Rio Grande do Sul. Disponível em: <<https://lume.ufrgs.br/handle/10183/277000>>.
- CORDEIRO, G. M.; DEMÉTRIO, C. G. Modelos lineares generalizados e extensões. *Piracicaba: USP*, 2008.
- DIXON, M. J.; COLES, S. G. Modelling association football scores and inefficiencies in the football betting market. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, Wiley, v. 46, n. 2, p. 265–280, 1997.
- KARLIS, D.; NTZOUFRAS, I. Modelling and simulating count data in r. *Journal of Statistical Software*, v. 14, n. 10, 2005. Disponível em: <<https://www.jstatsoft.org/article/view/v014i10>>.
- MAHER, M. J. Modelling association football scores. *Applied Statistics*, Wiley, v. 31, n. 2, p. 109–118, 1982.
- MONTGOMERY, D. C.; PECK, E. A.; VINING, G. G. *Introduction to linear regression analysis*. [S.l.]: John Wiley & Sons, 2021.
- PAULA, G. A. *Modelos de Regressão: Modelos Lineares, Poisson e Bivariados*. [S.l.]: IME-USP, 2024.
- R-BLOGGERS. *An Introduction to Modelling Soccer Matches in R: Part 1*. 2019. Disponível em: <https://www.r-bloggers.com/2019/05/an-introduction-to-modelling-soccer-matches-in-r-part-1/#google_vignette>.
- R-BLOGGERS. *An Introduction to Modelling Soccer Matches in R: Part 2*. 2020. Disponível em: <https://www.r-bloggers.com/2020/04/an-introduction-to-modelling-soccer-matches-in-r-part-2/#google_vignette>.
- SILVA, W. B. da. *Distribuição de Poisson Bivariada aplicada à previsão de resultados esportivos*. Tese (Doutorado) — Universidade Federal de São Carlos (UFSCar), 2014. Disponível em: <<https://repositorio.ufscar.br/bitstream/handle/ufscar/4586/6128.pdf?sequence=1>>.

Apêndice

```
library(rvest)
library(dplyr)
library(tidyr)
library(xml2)
library(tidyverse)
library(ggplot2)
library(worldfootballR)
library(purrr)
library(reshape2)
library(devtools)
library(MASS)

devtools::install_github("JaseZiv/worldfootballR")

# Definindo o país e a temporada

country <- "BRA"
season_end_year <- 2024

# Obtendo os resultados das partidas

brasileirao_matches <- fb_match_results(country = country, gender = "M",
season_end_year = season_end_year, tier = "1st")

# Convertendo os nomes dos times para fator

brasileirao_matches <- brasileiro_matches %>%
  mutate(
    Home = factor(Home),
    Away = factor(Away)
  )

# Preparando a base com resultado categórico
```

```
df_resultados <- brasileirao_matches %>%
  filter(!is.na(HomeGoals)) %>%
  mutate(
    Resultado = case_when(
      HomeGoals > AwayGoals ~ "Mandante",
      HomeGoals < AwayGoals ~ "Visitante",
      TRUE ~ "Empate"
    )
  )
```

Função para "derreter" os dados de rodadas (obter mandante e visitante em linhas separadas)

```
melt_results <- function(brasileirao_matches) {
  bind_rows(
    brasileirao_matches %>%
      transmute(
        team = Home,
        opponent = Away,
        location = "Home",
        g_for = HomeGoals,
        g_ag = AwayGoals
      ),
    brasileirao_matches %>%
      transmute(
        team = Away,
        opponent = Home,
        location = "Away",
        g_for = AwayGoals,
        g_ag = HomeGoals
      )
  )
}
```

```
# Calculando a média de gols marcados por localização

real_data_means <- brasileiro_matches %>%
  filter(!is.na(HomeGoals), !is.na(AwayGoals)) %>%
  # reorganiza os dados no formato longo
  select(Home, Away, HomeGoals, AwayGoals) %>%
  melt_results() %>%
  group_by(location) %>%
  summarise(mean_scored = mean(g_for))

real_data_means

simulated_poisson <- real_data_means %>%
  split(f = .$location) %>%
  map_df(function(x) {
    data.frame(
      dist = rpois(100000, lambda = x$mean_scored),
      location = x$location,
      data = "simulated"
    )
  })

# Gráfico da Distribuição de Gols Observados vs. Simulados

observed_data <- brasileiro_matches %>%
  filter(!is.na(HomeGoals), !is.na(AwayGoals)) %>%
  select(Home, Away, HomeGoals, AwayGoals) %>%
  melt_results() %>%
  select(location, dist = g_for) %>%
  mutate(data = "observed")

# Unindo com os dados simulados
```

```
dados_completos <- bind_rows(simulated_poisson, observed_data)

dados_completos$data <- factor(dados_completos$data,
levels = c("observed", "simulated"))

dados_completos <- dados_completos %>%
  mutate(
    location = tolower(location),
    data = case_when(
      data == "observed" ~ "observed",
      data == "simulated" ~ "simulated",
      TRUE ~ data
    )
  )

ggplot(dados_completos, aes(x = dist, fill = data)) +
  geom_bar(position = "dodge", aes(y = ..prop.., group = data), stat = "count",
  color = "black") +
  facet_wrap(
    ~location,
    labeller = as_labeller(c(home = "Mandante", away = "Visitante"))
  ) +
  scale_fill_manual(
    values = c("observed" = "red", "simulated" = "blue"),
    labels = c("observed" = "Observado", "simulated" = "Simulado")
  ) +
  labs(
    title = "Distribuição de Gols Observados vs. Simulados",
    x = "Número de Gols",
    y = "Proporção",
    fill = "Tipo de Dado"
  ) +
  theme_minimal()
```

```
calc_chi_squared <- function(location, brasileirao_matches) {
  if (!location %in% c("home", "away")) {
    stop("game_location deve ser 'home' ou 'away'")
  }

  # Gerar dados longos
  dados_longos <- brasileirao_matches %>%
    filter(!is.na(HomeGoals), !is.na(AwayGoals)) %>%
    select(Home, Away, HomeGoals, AwayGoals) %>%
    melt_results()

  # Filtrar gols feitos pela localização desejada
  goals_scored <- dados_longos %>%
    filter(location == location) %>%
    pull(g_for)

  # Verifica se há gols observados
  if (length(goals_scored) == 0 || sum(goals_scored) == 0) {
    stop(paste("Sem gols observados para localização:", game_location))
  }

  # Frequência observada
  observed_goal_counts <- table(goals_scored)

  # Média de gols
  mean_goals <- mean(goals_scored)

  # Calcula as probabilidades de Poisson
  max_goals <- max(goals_scored)
  probs <- dpois(0:max_goals, lambda = mean_goals)
  probs <- append(probs, 1 - sum(probs))

  # Frequência observada ajustada (para ter mesmo comprimento que probs)
  freq_vector <- as.numeric(observed_goal_counts)
  freq_vector <- c(freq_vector, 0) # adiciona zero para a cauda
```

```
# Teste qui-quadrado
test <- chisq.test(x = freq_vector, p = probs, simulate.p.value = TRUE)

return(test)
}

# Valores de qui-quadrado

calc_chi_squared("home", brasileirao_matches)
calc_chi_squared("away", brasileirao_matches)


library(bivpois)

load("C:/Users/Fraga/Desktop/bivpois.RData")


# Modelando a força ofensiva e defensiva dos times com efeitos fixos:

# Fórmula para modelar lambda1 (mandante) e lambda2 (visitante)

form1 <- ~c(Home, Away) + c(Away, Home)

# Modelo 1 { Double Poisson (sem dependência)

ex4.m1 <- lm.bp(HomeGoals~1, AwayGoals~1, l1l2=form1,
zeroL3=TRUE, data=brasileirao_matches)

# Modelo 2 { Bivariado com constante

ex4.m2 <- lm.bp(HomeGoals~1, AwayGoals~1,
l1l2 = form1, data = brasileirao_matches)

# Modelo 3 { depende do time da casa
```

```
ex4.m3 <- lm.bp(HomeGoals~1, AwayGoals~1, l1l2 = form1,  
l3 = ~Home, data = brasileiro_matches)
```

```
# Modelo 4 { depende do time visitante
```

```
ex4.m4 <- lm.bp(HomeGoals~1, AwayGoals~1, l1l2 = form1, l3 = ~Away,  
data = brasileiro_matches)
```

```
# Modelo 5 { depende de ambos os times
```

```
ex4.m5 <- lm.bp(HomeGoals~1, AwayGoals~1, l1l2 = form1,  
l3 = ~Home + Away, data = brasileiro_matches)
```

```
# Modelo 6 { Zero-inflated (Poisson Bivariado com excesso de zeros)
```

```
ex4.m6 <- lm.dibp(HomeGoals~1, AwayGoals~1,  
l1l2=form1, data=brasileirao_matches, jmax=0)
```

```
# Modelo 7 { Distribuição geométrica para empates (  
A distribuição geométrica favorece mais empates 0x0 e decai rapidamente.)
```

```
ex4.m7 <- lm.dibp(HomeGoals~1, AwayGoals~1,  
l1l2=form1, data=brasileirao_matches, distribution="geometric"
```

```
# Modelos com inflação até j gols (empates 0x0, 1x1, etc.)
```

```
ex4.m8 <- lm.dibp(HomeGoals~1, AwayGoals~1,  
l1l2=form1, data=brasileirao_matches, jmax=1)
```

```
ex4.m9 <- lm.dibp(HomeGoals~1, AwayGoals~1,  
l1l2=form1, data=brasileirao_matches, jmax=2)
```

```
ex4.m10 <- lm.dibp(HomeGoals~1, AwayGoals~1,  
l1l2=form1, data=brasileirao_matches, jmax=3)
```

```
# Modelo com inflação de empates via Poisson
```

```
(A Poisson distribui a inflação ao longo de empates com mais suavidade que a geométrica.)
```

```
ex4.m11 <- lm.dibp(HomeGoals~1, AwayGoals~1,
l1l2=form1, data=brasileirao_matches, distribution="poisson")

# Modelo 12 { Versão Double Poisson com inflação na diagonal

ex4.m12 <- lm.dibp(HomeGoals~1, AwayGoals~1, l1l2=form1, data=brasileirao_matches,
                  zeroL3=TRUE)

## Função ataque e defesa

# Transformando em formato longo com melt_results()
melt_results <- function(dados) {
  casa <- dados %>%
    transmute(team = Home,
              location = "home",
              g_for = HomeGoals,
              g_ag = AwayGoals)

  fora <- dados %>%
    transmute(team = Away,
              location = "away",
              g_for = AwayGoals,
              g_ag = HomeGoals)

  bind_rows(casa, fora)
}

# Aplicando a função para calcular ataque (alpha) e defesa (beta)

basic_model <- brasileiro_matches %>%
  filter(!is.na(HomeGoals), !is.na(AwayGoals)) %>%
  melt_results() %>%
  group_by(team) %>%
```

```
summarise(
  alpha = mean(g_for),
  beta = mean(g_ag)
) %>%
arrange(desc(alpha))

# Extraíndo os coeficientes do modelo M12

coefs <- coef(ex4.m12)

# Extraíndo apenas os coeficientes de ataque (11) e defesa (12)

alpha_coefs <- coefs[grepl("^\\(11\\):Away\\.\\.\\.Home", names(coefs))]
beta_coefs <- coefs[grepl("^\\(12\\):Home\\.\\.\\.Away", names(coefs))]

# Limpando os nomes

limpar_nomes <- function(nomes) {
  nomes %>%
    gsub("^\\(1[12]\\):Away\\.\\.\\.Home", "", .) %>%
    gsub("^\\(1[12]\\):Home\\.\\.\\.Away", "", .) %>%
    trimws()
}

times_alpha <- limpar_nomes(names(alpha_coefs))
times_beta <- limpar_nomes(names(beta_coefs))

# Conferindo se estão no mesmo padrão

stopifnot(identical(sort(times_alpha), sort(times_beta)))

# Montando o data frame final

forcas_df <- data.frame(
  Time = times_alpha,
  'Força Ofensiva ()' = round(-alpha_coefs, 3),
```

```
'Força Defensiva ()' = round(-beta_coefs, 3)
)

colnames(forcas_df) <- c("Time", "Força Ofensiva ()", "Força Defensiva ()")

forcas_df <- forcas_df %>%
  dplyr::arrange(dplyr::desc('Força Ofensiva ()))

print(forcas_df)

# Reestruturando dados para gráfico
forcas_long <- forcas_df %>%
  pivot_longer(cols = c('Força Ofensiva ()', 'Força Defensiva ()'),
               names_to = "Tipo",
               values_to = "Valor")

# Gerando o gráfico de Forças Ofensivas e Defensivas por Time

ggplot(forcas_long, aes(x = reorder(Time, Valor), y = Valor, fill = Tipo)) +
  geom_bar(stat = "identity", position = "dodge") +
  coord_flip() +
  labs(x = "Times", y = "Valor estimado",
       title = "Forças Ofensivas e Defensivas por Time") +
  scale_fill_manual(values = c(
    "Força Ofensiva ()" = "steelblue", "Força Defensiva ()" = "firebrick")) +
  theme_minimal()

# Função para gerar a tabela do campeonato

results_to_table <- function(brasileirao_matches) {
  melt_results(brasileirao_matches) %>%
    mutate(
      points = case_when(
```

```

      g_for > g_ag ~ 3,
      g_for == g_ag ~ 1,
      TRUE ~ 0
    ),
    gd = g_for - g_ag
  ) %>%
group_by(team) %>%
summarise(
  Partidas = n(),
  GolsPro = sum(g_for),
  GolsContra = sum(g_ag),
  Saldo = sum(gd),
  Pontos = sum(points),
  .groups = 'drop'
) %>%
  arrange(desc(Pontos), Saldo, GolsContra)
}

# Incorporando o banco

league_table <- brasileirao_matches %>%
  filter(!is.na(HomeGoals) & !is.na(AwayGoals)) %>%
  results_to_table()

print(league_table)

# Gráfico com os resultados dos confrontos

p1 <- ggplot(df_resultados, aes(x = Away, y = Home, fill = Resultado)) +
  geom_tile(color = "black", linewidth = 0.25) +
  geom_label(aes(label =
    paste(HomeGoals, AwayGoals, sep = "-")), fill = "white", size = 3) +
  scale_fill_manual(values = c("Mandante" = "green",
    "Visitante" = "red", "Empate" = "gray80")) +

```

```
scale_x_discrete(position = "top") +
scale_y_discrete(limits = rev(levels(factor(brasileirao_matches$Home)))) +
labs(
  x = "Time Visitante",
  y = "Time Mandante",
  title = "Resultados dos Confrontos - Brasileirão 2024",
  subtitle = "Verde = vitória do mandante |
  Vermelho = vitória do visitante | Cinza = empate"
) +
theme_minimal() +
theme(
  axis.text.x = element_text(angle = 45, hjust = 0, vjust = 1),
  axis.text = element_text(size = 9),
  plot.title = element_text(size = 14, face = "bold"),
  plot.subtitle = element_text(size = 10),
  panel.grid = element_blank()
)

p1
ggsave("Confrontos.png", plot = p1, width = 10, height = 8, dpi = 300)

# Analisando os modelos

modelos <- list(ex4.m1, ex4.m2, ex4.m3, ex4.m4, ex4.m5, ex4.m6, ex4.m7, ex4.m8,
ex4.m9, ex4.m10, ex4.m11, ex4.m12)
nomes <- paste0("M", 1:12)

# Extraíndo AIC e BIC diretamente dos modelos

aic_values <- sapply(modelos, function(m) m$AIC);aic_values
bic_values <- sapply(modelos, function(m) m$BIC);bic_values

# Gráfico de comparação de modelos
```

```
# Dados
modelos_labels <- paste0("M", 1:12)
aic_vals <- c(2103.266, 2105.269, 2137.920, 2134.752, 2151.643, 2107.269,
              2109.269, 2108.060, 2105.321, 2107.321, 2109.268, 2107.267)
bic_vals <- c(2288.599, 2295.235, 2415.919, 2412.751, 2517.675, 2301.868,
              2308.501, 2307.292, 2309.187, 2315.820, 2308.501, 2301.866)

# Data frame
df_modelos <- data.frame(
  Modelo = factor(modelos_labels, levels = modelos_labels),
  AIC = aic_vals,
  BIC = bic_vals
)

df_modelos_long <- pivot_longer(df_modelos,
  cols = c("AIC", "BIC"), names_to = "Critério", values_to = "Valor")

# Visualizando o gráfico

ggplot(df_modelos_long, aes(x = Modelo, y = Valor, color = Critério, group = Critério))
  geom_line(size = 1) +
  geom_point(size = 2.5) +
  labs(title = "Comparação dos Modelos M1 a M12",
        subtitle = "Critérios AIC e BIC",
        y = "Valor do Critério",
        x = "Modelo") +
  theme_minimal()

ggsave("comparacao_modelos.png", width = 8, height = 5)

# Ajustando o modelo de Poisson
modelo_poisson <- glm(gols ~ home + team + opponent,
  data = dados_long,
```

```
family = poisson(link = "log"))

# Função para prever gols e probabilidades de resultado

prever_prob_resultado <- function(time_casa, time_fora,
modelo = modelo_poisson, max_gols = 5) {
  team_levels <- levels(modelo$data$team)

  if (!(time_casa %in% team_levels) || !(time_fora %in% team_levels)) {
    return(NULL)
  }

  data_pred <- data.frame(
    home = c(1, 0),
    team = factor(c(time_casa, time_fora), levels = team_levels),
    opponent = factor(c(time_fora, time_casa), levels = team_levels)
  )

  mean_goals <- exp(predict(modelo, newdata = data_pred))

  placares <- expand.grid(
    gols_casa = 0:max_gols,
    gols_fora = 0:max_gols
  )

  placares$prob <- dpois(placares$gols_casa, lambda = mean_goals[1]) *
    dpois(placares$gols_fora, lambda = mean_goals[2])

  prob_casa <- sum(placares$prob[placares$gols_casa > placares$gols_fora])
  prob_empate <- sum(placares$prob[placares$gols_casa == placares$gols_fora])
  prob_fora <- sum(placares$prob[placares$gols_casa < placares$gols_fora])

  tibble::tibble(
    Casa = time_casa,
```

```
    Fora = time_fora,
    Gols_Esp_Casa = round(mean_goals[1], 2),
    Gols_Esp_Fora = round(mean_goals[2], 2),
    Prob_Casa = round(prob_casa, 3),
    Prob_Empate = round(prob_empate, 3),
    Prob_Fora = round(prob_fora, 3)
  )
}

# Lista dos times

times <- levels(modelo_poisson$data$team)

# Expandindo todos os confrontos diferentes

confrontos <- expand.grid(Casa = times, Fora = times, stringsAsFactors = FALSE) %>%
  filter(Casa != Fora)

# Aplicando a função a todos os confrontos

tabela_confrontos <- purrr::pmap_dfr(
  list(confrontos$Casa, confrontos$Fora),
  prever_prob_resultado
)

tabela_confrontos

# Média, moda e mediana dos gols esperados em todos os confrontos

# Função para calcular a moda
moda <- function(x) {
  ux <- unique(x)
  freq <- tabulate(match(x, ux))
  ux[which.max(freq)]
}
```

```
# Unindo os gols esperados em um único vetor
gols_esperados <- c(tabela_confrontos$Gols_Esp_Casa,
tabela_confrontos$Gols_Esp_Fora)

# Cálculos por tipo
media_mandante <- mean(tabela_confrontos$Gols_Esp_Casa)
mediana_mandante <- median(tabela_confrontos$Gols_Esp_Casa)
moda_mandante <- moda(round(tabela_confrontos$Gols_Esp_Casa, 2))

media_visitante <- mean(tabela_confrontos$Gols_Esp_Fora)
mediana_visitante <- median(tabela_confrontos$Gols_Esp_Fora)
moda_visitante <- moda(round(tabela_confrontos$Gols_Esp_Fora, 2))

# Resultados

cat("Média:", round(media_mandante, 3), "\n")
cat("Mediana:", round(mediana_mandante, 3), "\n")
cat("Moda:", moda_mandante, "\n\n")

cat("Média:", round(media_visitante, 3), "\n")
cat("Mediana:", round(mediana_visitante, 3), "\n")
cat("Moda:", moda_visitante, "\n")

# Filtrando apenas os confrontos da 1ª rodada

rodada_1 <- tabela_confrontos %>%
  filter(
    (Casa == "Vitória" & Fora == "Palmeiras") |
    (Casa == "Cruzeiro" & Fora == "Botafogo (RJ)") |
    (Casa == "Atl Goianiense" & Fora == "Flamengo") |
    (Casa == "Corinthians" & Fora == "Atlético Mineiro") |
    (Casa == "Vasco da Gama" & Fora == "Grêmio") |
    (Casa == "Ath Paranaense" & Fora == "Cuiabá") |
    (Casa == "Fluminense" & Fora == "RB Bragantino") |
```

```

      (Casa == "São Paulo" & Fora == "Fortaleza") |
      (Casa == "Internacional" & Fora == "Bahia") |
      (Casa == "Criciúma" & Fora == "Juventude")
    ) %>%
  mutate(
    Confronto = paste(Casa, "x", Fora),
    Confronto = fct_reorder(Confronto, Prob_Casa)
  )

# Reorganizando em formato longo e renomeando os resultados

df_long <- rodada_1 %>%
  pivot_longer(cols = c(Prob_Casa, Prob_Empate, Prob_Fora),
               names_to = "Resultado",
               values_to = "Probabilidade") %>%
  mutate(
    Resultado = recode(Resultado,
                       Prob_Casa = "Vitória Mandante",
                       Prob_Empate = "Empate",
                       Prob_Fora = "Vitória Visitante"),
    Resultado = factor(Resultado, levels = c(
      "Vitória Mandante", "Empate", "Vitória Visitante"))
  )

# Gráfico de probabilidades previstas

ggplot(df_long, aes(x = Probabilidade, y = fct_rev(factor(Confronto)), fill = Resultado)) +
  geom_bar(stat = "identity", position = position_stack(reverse = TRUE)) +
  geom_text(aes(label = paste0(round(100 * Probabilidade), "%")),
            position = position_stack(vjust = 0.5, reverse = TRUE),
            size = 3.5, color = "white") +
  scale_fill_manual(
    values = c(
      "Vitória Mandante" = "green3",
      "Empate" = "gray60",
      "Vitória Visitante" = "red2"
    )
  )

```

```
)  
) +  
labs(  
  title = "Probabilidades Previstas { 1ª Rodada do Brasileirão",  
  x = "Probabilidade",  
  y = "Confronto",  
  fill = "Resultado Previsto"  
) +  
theme_minimal(base_size = 12) +  
theme(  
  plot.title = element_text(hjust = 0.5, face = "bold"),  
  legend.position = "bottom"  
)  
  
ggsave("previsões.png", width = 9, height = 6, dpi = 300)
```