Revista Brasileira de Futsal e Futebol
ISSN 1984-4956 ***versão eletrônica***

Periódico do Instituto Brasileiro de Pesquisa e Ensino em Fisiologia do Exercício

w w w . i b p e f e x . c o m . b r / w w w . r b f f . c o m . b r

CORREÇÃO DE EMPATES PARA MODELAGEM DINÂMICA DE PARTIDAS DE FUTEBOL

Anderson Ribeiro Duarte[1]
Helgem de Souza Ribeiro Martins[2]
Bruno Fernandes da Silva[1]

RESUMO

O futebol está entre os esportes de mais difícil previsibilidade, ou seja, no qual a ocorrência de resultados atípicos, em que equipes inferiores suplantam as equipes melhores se torna quase corriqueiro. Esse trabalho apresenta um mecanismo já existente para se obter previsibilidade para o resultado de partidas de futebol através de um modelo Poisson truncado à direita. Além disso apresenta uma nova proposta de adaptação da cota dinâmica para a determinação de empates já utilizada no modelo anterior. O modelo de simulação utilizado é descrito assim como as estratégias para produção das cotas de correção de empates e resultados promissores são discutidos para o Campeonato Brasileiro de Futebol da Série A em 2013 e 2015.

Palavras-chave: Futebol. Poisson truncada. Campeonato Brasileiro de Futebol. Modelo linear generalizado Beta.

ABSTRACT

Tie correction for dynamic modeling of football matches

Football is among the most difficult predictability sports, in which the occurrence of atypical results, in which inferior teams supplant the better teams becomes almost commonplace. This work presents an already existent mechanism to obtain predictability to outcome of matches result through a truncated Poisson. In addition, it presents a novel proposal for adapting the dynamic threshold to determination ties in the previous model. The simulation model used is described and also the strategies for producing the correct tie threshold and promising results are discussed for the Brazilian Serie A Football Championship in 2013 and 2015.

Key words: Football. Truncated Poisson. Brazilian Football Championship. Generalized linear model Beta.

E-mails dos autores:
anderson.duarte@ufop.edu.br
helgem@ufop.edu.br
bruno.fernandes@aluno.ufop.edu.br

1-Departamento de Estatística, Universidade Federal de Ouro Preto (UFOP), Campus Morro do Cruzeiro, Ouro Preto-MG, Brasil.
2-Pro-Reitoria de Pesquisa e Pós-Graduação, Universidade Federal de Ouro Preto (UFOP) Campus Morro do Cruzeiro, Ouro Preto-MG, Brasil.

Endereço para correspondência:
Anderson Ribeiro Duarte
Departamento de Estatística, Universidade Federal de Ouro Preto (UFOP), Campus Morro do Cruzeiro, Ouro Preto-MG, Brasil.
CEP: 35400-000

Revista Brasileira de Futsal e Futebol
ISSN 1984-4956 ***versão eletrônica***

Periódico do Instituto Brasileiro de Pesquisa e Ensino em Fisiologia do Exercício

w w w . i b p e f e x . c o m . b r / w w w . r b f f . c o m . b r

INTRODUÇÃO

Dentre todos os esportes de prática profissional no mundo, o futebol tem papel de grande destaque. Trata-se de um esporte com um número de praticantes extremamente vultoso. Os números são também muito relevantes quando consideramos os que não são praticantes profissionais, mas se incluem nesse meio como aficionados acompanhantes desse esporte clássico. A relevância pode ser notada por meio de grande espaço cedido em todos os veículos de mídia.

Em diversas localidades, o envolvimento com o futebol em nível profissional movimenta somas financeiras absurdas, e obviamente, sistemas de apostadores (legalizados em alguns países e ilegais em outros) também movimentam valores expressivos. Busca-se o acerto de resultados de partidas, placares exatos entre outras informações específicas do jogo.

Dadas os relevantes valores financeiros nas principais equipes ao longo do mundo, em uma análise ingênua, seria previsível associar a chance de vitória de uma equipe ao seu poderio financeiro. Por outro lado, resultados diversos funcionam como um claro contraponto para tal evidência.

Esportes de baixa pontuação são de difícil previsibilidade para seus resultados. Nesse cenário, a ocorrência de resultados atípicos é recorrente, ou seja, resultados em que equipes inferiores superam as equipes melhores ocorrem com certa frequência. Esse fenômeno é mais comum em esportes coletivos e mais ainda nos casos de esportes de pontuação baixa. Por exemplo, resultados atípicos são menos frequentes em esportes como o voleibol quando comparado ao futebol.

Este conjunto de fatores é suficiente para transformar a tarefa de produção de modelos para previsibilidade dos resultados do jogo como algo de extrema dificuldade. O meio esportivo tem se tornado alvo da discussão de diversos especialistas na inserção de modelos matemáticos e estatísticos com interesse em explicar e fornecer previsibilidade para os resultados. Diversos trabalhos vêm sendo desenvolvidos, não somente com o futebol, mas com outros esportes também, como por exemplo em Knorr-Held (2000), Rue e Salvesen (2000) e Souza Jr. e Gamerman (2004).

O avanço tecnológico trouxe grande avanço na capacidade de manipulação de dados computacionalmente. Aliado a este fato, as técnicas estatísticas também experimentaram uma grande evolução nos últimos anos. Estes dois fatores, em conjunto, possibilitam um melhor tratamento do problema foco deste estudo. O trabalho de produzir modelos para previsibilidade de resultados de partidas de futebol pode ser observado como um processo de decisão para escolhas de investimento em equipes e campeonatos. Algum tipo de controle sobre os resultados das partidas de futebol é na prática algo de valor inestimável.

Mais que jogar futebol ou torcer para alguma agremiação esportiva, há um expressivo interesse em prever o resultado de um jogo ou a classificação final de um campeonato. A motivação deste objetivo é claramente financeira, devido principalmente às cifras vultuosas envolvidas no esporte. Os estudos clássicos, em geral, utilizam séries históricas com intervalo temporal curto para analisar possíveis resultados para partidas ainda não realizadas.

Autores como Faria (2005) e Macedo e Silva (2014), tentam, através de modelagem estatística, prever o resultado de um jogo e mesmo de um campeonato. Martins e Duarte (2014) propõem um modelo que utiliza-se de uma distribuição Poisson truncada à direita, com valor esperado sendo atualizado a cada rodada, baseado no desempenho dos clubes, para predizer o resultado de determinada partida.

O trabalho apresentado por Martins e Duarte (2014) deixa claro um interesse central em estimar resultados de partidas, ou seja, não trata de uma metodologia de previsão para campeonatos. Uma grande contribuição verificada nesse trabalho é uma medida de correção para previsão de resultados de empate, dado que diversos modelos da literatura mostram deficiência neste fato. Martins e Duarte (2014) já apresentam uma medida de correção obtida empiricamente por um volume grande de testes, porém não deixando uma clara justificativa para sua adoção. O presente trabalho tem por finalidade propor uma estratégia de correção de empates que seja totalmente dependente dos dados, utilizando-se de algum modelo estatístico.

Esse texto se encontra organizado da seguinte forma: a segundo seção apresenta

Revista Brasileira de Futsal e Futebol
ISSN 1984-4956 ***versão eletrônica***

Periódico do Instituto Brasileiro de Pesquisa e Ensino em Fisiologia do Exercício

w w w . i b p e f e x . c o m . b r / w w w . r b f f . c o m . b r

uma revisão bibliográfica de alguns métodos nessa área de estudo, além de um detalhamento da estrutura de modelagem dinâmica para partidas de futebol proposta por Martins e Duarte (2014); a terceira seção faz uma análise descritiva acerca de dados com resultados de partidas no Campeonato Brasileiro de futebol Série A de 2015 e apresenta a nova proposta para estimativa do percentual de empates em rodadas a serem simuladas, através de um modelo linear generalizado Beta; a quarta seção descreve o conjunto de experimentos realizado, contemplando o Campeonato Brasileiro da série A nos anos de 2013 e 2015 atestando a qualidade das propostas apresentadas; a quinta seção discute os resultados alcançados e por fim, a seção final apresenta considerações acerca desse estudo e propõe alguns caminhos de continuidade de estudos.

Revisão de Literatura

A literatura acerca deste assunto é um tanto escassa. Entretanto algumas discussões sobre modelos de previsibilidade para futebol podem ser encontradas. Knorr-Held (2000) considera o problema de classificar dinamicamente equipes esportivas com base em resultados categóricos de comparações pareadas, como vitória, empate e perda no futebol. Uma estrutura de modelagem via modelo de link cumulativo para respostas ordenadas, em que os parâmetros latentes representam a força de cada equipe. Uma extensão dinâmica deste modelo é proposta com conexões próximas aos métodos de suavização não paramétricos.

Karlis e Ntzoufras (2000) avaliam pressupostos relevantes na discussão do assunto, como por exemplo a possível independência entre gols anotados por oponentes em uma partida e propondo que o vetor bivariado de gols feitos pelo mandante e gol feitos pelo visitante em uma partida segue uma distribuição bivariada de Poisson, independente do parâmetro de correlação. O estudo avalia a viabilidade da utilização dessa distribuição em modelos discretos para simulação de resultados de partidas.

Posteriormente, Karlis e Ntzoufras (2003) exploram e aplicam modelagens para alguns esportes coletivos, como futebol e polo aquático. De modo geral, os modelos propõem uma abordagem via distribuição de Poisson, com taxa de distribuição baseada em diversos fatores, tais como número de gols como mandante, número de gols sofridos pelo adversário, mando de campo, dentre outros. Um modelo é proposto para prever o resultado de um jogo de futebol dos resultados anteriores de ambas as equipes (Rotshtein, Posner e Rakityanskaya, 2005).

Este modelo é subjacente ao método de identificação de dependências não-lineares por bases de conhecimento difusas. Os resultados aceitáveis da simulação podem ser obtidos ajustando as regras fuzzy usando os dados do torneio. O procedimento de ajuste implica a escolha dos parâmetros de funções de associação de termo difuso e pesos de regra por uma combinação de técnicas de otimização genética e neural.

Liu e Zhang (2008) discutem um modelo de regressão polinomial e Brillinger (2009) constrói um modelo trinomial em que os três resultados possíveis para o modelo trinomial contemplam os três resultados possíveis em uma partida de futebol (vitória do mandante, empate e vitória do visitante). Faria (2005) e também Souza Jr. e Gamerman (2004) seguem uma abordagem bayesiana, cujo interesse é modelar o número de gols através de uma distribuição Poisson e possíveis variações.

Alves e colaboradores (2011) aplicam dois modelos logit ordinais para ajustar os resultados dos jogos no futebol brasileiro. Como variáveis explicativas são empregadas medidas de desempenho anterior das equipes ao longo de todos os jogos anteriores, ao longo de jogos recentes e ao jogar em casa e como visitante. Os resultados do ajuste dos modelos são empregados nas simulações realizadas para prever o número de pontos a serem ganhos nos jogos a seguir e antecipar a classificação final das equipes.

Martins e Duarte (2014) buscam obter previsibilidade para o resultado de partidas de futebol através de um modelo Poisson truncado à direita e uma cota teoricamente dinâmica para a determinação de resultados de empates. Esse trabalho é o background para a proposição do presente estudo. Dessa forma, uma descrição mais detalhada sobre as discussões de Martins e Duarte (2014) será necessária.

Revista Brasileira de Futsal e Futebol
ISSN 1984-4956 ***versão eletrônica***

Periódico do Instituto Brasileiro de Pesquisa e Ensino em Fisiologia do Exercício

w w w . i b p e f e x . c o m . b r / w w w . r b f f . c o m . b r

MATERIAIS E MÉTODOS

Martins e Duarte (2014) fazem simulações de forma que, ao final destas, seja apresentada a probabilidade de vitória do time mandante, ora nominada Vm, a probabilidade de que ocorra um empate, denominada E e, por fim, a probabilidade de que ocorra a vitória do time visitante, chamada Vv. Num primeiro momento, os autores representam o resultado de forma vetorial através de $(1,0,0)$ para indicar a vitória do time MANDANTE, $(0,1,0)$ para representar um EMPATE, e $(0,0,1)$ para representar uma vitória do time VISITANTE.

A proposta inicial dos autores para um estimador do resultado da partida foi dada por:

$$\hat{\theta} = \begin{cases} (1,0,0), & \text{se } \max(\mathrm{Vm};\ \mathrm{E};\ \mathrm{Vv}) = \mathrm{Vm}; \\ (0,1,0), & \text{se } \max(\mathrm{Vm};\ \mathrm{E};\ \mathrm{Vv}) = \mathrm{E}; \\ (0,0,1), & \text{se } \max(\mathrm{Vm};\ \mathrm{E};\ \mathrm{Vv}) = \mathrm{Vv}; \end{cases}$$

tal critério se mostrou adequado na previsão da vitória do time visitante, entretanto era deficitário em prever resultados de empates, transformando, em muitos casos, resultados que deveriam ser previstos como empates em vitórias do time mandante. Um segundo estimador foi proposto, a correção se baseou em uma cota específica que limitava a estimação em resultados de vitórias de mandantes. Seja φ o percentual de empates ocorridos até a rodada imediatamente anterior à rodada da previsão que será realizada. A cota proposta é dada por $\text{cota} = 1{,}07\varphi$, assim o estimador passou ao seguinte formato:

$$\hat{\theta} = \begin{cases} (1,0,0), & \text{se } \max(\mathrm{Vm};\ \mathrm{E};\ \mathrm{Vv}) = \mathrm{V_m} \text{ e } \mathrm{E} < \text{cota}; \\ (0,1,0), & \text{se } \max(\mathrm{Vm};\ \mathrm{E};\ \mathrm{Vv}) = \mathrm{V_m} \text{ e } \mathrm{E} \geq \text{cota ou } \max(\mathrm{Vm};\ \mathrm{E};\ \mathrm{Vv}) = \mathrm{E}; \\ (0,0,1), & \text{se } \max(\mathrm{Vm};\ \mathrm{E};\ \mathrm{Vv}) = \mathrm{Vv}; \end{cases}$$

essa transformação fez com que as previsões do modelo se mostrassem mais compatíveis com os resultados reais das partidas. Entretanto, estudos que objetivassem verificar se $1{,}07$ é o valor mais adequado para a constante utilizada para determinar o valor para a cota, ou mesmo se esta constante deve variar de campeonato para campeonato, ou ainda se deve variar de rodada a rodada não foram conduzidos. Ademais, seria interessante que uma análise descritiva dos resultados simulados fosse feita e comparada com a análise descritiva dos resultados reais.

Distribuição de Poisson Truncada à direita

Ao se verificar o histórico de placares em torneios de futebol ao redor do globo, nota-se que são raras as ocorrências de placares extremamente dilatados, com diferenças entre as equipes sendo superiores à 6 gols por exemplo. Ao se definir uma taxa para a distribuição dos gols, utilizando-se a distribuição clássica de Poisson na simulação de resultados, existe a possibilidade de ocorrência de placares simulados superestimados. Para redução da ocorrência de tais placares, foi proposto fixar uma cota superior à distribuição dos gols marcados por cada time. Desta forma, surge o modelo de Poisson truncado à direita para simulação, que limita o espaço paramétrico dos gols a serem marcados pelos competidores. Seja X o número de gols marcados por uma equipe em uma determinada partida e seja $\mathrm{x_{max}}$ a cota superior máxima de gols que tal equipe pode fazer na referida partida. As probabilidades para a variável aleatória X podem ser obtidas por:

$$\mathrm{P}(\mathrm{X} = \mathrm{x} | \mathrm{X} \leq \mathrm{x_{max}}) = \frac{{}^{\mathrm{e^{-\lambda}\lambda^{x}}}/_{\mathrm{x!}}}{\sum_{\mathrm{k=0}}^{\mathrm{x_{max}}} {}^{\mathrm{e^{-\lambda}\lambda^{k}}}/_{\mathrm{k!}}}$$

em que λ representa a taxa da distribuição Poisson, as taxas utilizadas para a equipe mandante λ_{m}^{+}, e para a equipe visitante λ_{v}^{+} são obtidas através de informações das rodadas anteriormente disputadas com fatores nominados fator ataque e fator defesa de cada equipe, como segue:

μ_{n}^{+} = média de gols feitos pelo time mandante nas últimas n rodadas como mandante;

μ_{n}^{-} = média de gols sofridos pelo time mandante nas últimas n rodadas como mandante;

γ_{n}^{+} = média de gols feitos pelo time visitante nas últimas n rodadas como mandante;

γ_{n}^{-} = média de gols sofridos pelo time visitante nas últimas n rodadas como mandante.

Revista Brasileira de Futsal e Futebol
ISSN 1984-4956 ***versão eletrônica***

Periódico do Instituto Brasileiro de Pesquisa e Ensino em Fisiologia do Exercício

w w w . i b p e f e x . c o m . b r / w w w . r b f f . c o m . b r

De posse de tais fatores, foram estabelecidas as seguintes taxas:

$$\lambda_m^+ = \max\left(1, \frac{\mu_n^+ + \gamma_n^-}{2}\right) \text{ e } \lambda_v^+ = \max\left(1, \frac{\mu_n^- + \gamma_n^+}{2}\right).$$

Note que o método utilizado por Martins e Duarte (2014) confronta a qualidade dos ataques com a resistência das defesas. É esperado equilíbrio entre um ataque que venha marcando muitos gols e uma defesa que sofra poucos, e tal método prevê, de certa forma, tal equilíbrio.

Em casos de taxa média inferiores a 1, a mesma não foi utilizada e aponta-se uma justificativa técnica para a escolha do valor 1. No caso de taxas pequenas, quando a taxa é inferior a 1, a probabilidade do valor 0 fica elevada, e este resultado será privilegiado, prejudicando, assim, as estimativas para a previsão do resultado do jogo.

O número de rodadas (n) consideradas da série histórica deve ser capaz de captar informações efetivas sobre a equipe. Neste o trabalho, o valor (n) será fixado em 5, pois a intenção do modelo é considerar intervalos relativamente curtos, de aproximadamente dois meses. Observa-se que em um período de dois meses cada time participa de aproximadamente 10 jogos em ligas nacionais em diversas localidades, sendo, em geral, cinco jogos em sua sede e outros cinco no estádio do adversário, frequentemente distribuídos de forma alternada.

Além disto, é preciso definir o ponto de truncamento x_{max}. Seja $\max G_m$ o número máximo de gols que a equipe mandante fez em um único jogo como mandante nas partidas consideradas, e $\max G_v$ o número máximo de gols que a equipe visitante fez como visitante nas partidas consideradas para a simulação. O ponto de truncamento x_{max} foi definido por $\max(\max G_m, \max G_v) + 1$, ou seja, o truncamento foi único para as duas equipes. Além disso, soma-se uma unidade ao valor máximo dos gols já feitos com o intuito de não limitar as equipes, nas rodadas seguintes, a um desempenho que seja no máximo idêntico àquele já obtido anteriormente, mas permite que ambas possam apresentar desempenho superior ao já demonstrado.

O modelo computacional de simulação é executado em duas etapas. Inicialmente são calculados as taxas e o ponto de truncamento para cada partida, então é construída uma distribuição bivariada $(gols_m, gols_v)$, em que $gols_m$ e $gols_v$ são duas variáveis aleatórias Poisson truncadas à direita em x_{max}, independentes e com médias λ_m^+ e λ_v^+, respectivamente.

A segunda etapa do modelo computacional utilizado por Martins e Duarte (2014) busca encontrar o vetor aleatório $(gols_m, gols_v)$ para diversas repetições e baseados na contagem de ocorrências $gols_m > gols_v$, $gols_m = gols_v$ e $gols_m < gols_v$ estimar as probabilidades Vm, E e Vv.

RESULTADOS

A base de dados utilizada foi constituída pelos resultados do Campeonato Brasileiro da série A em 2015. Existem razões específicas para a utilização dessa base de dados. Primeiramente, o Campeonato Brasileiro da série A é notoriamente visto como um campeonato de extremo equilíbrio entre as equipes, fazendo com que probabilidades previstas para empates, vitorias de mandantes e visitantes se tornem bastante equilibradas. Essa análise do equilíbrio é relevante visto que o alvo central desse trabalho é melhorar a condição de previsão do resultado de empate nas partidas. Em campeonatos de muito equilíbrio, se torna recorrente que em partidas com previsão de vitoria do mandante se transformem em resultado real de empate, ou vice-versa.

O modelo proposto por Martins e Duarte (2014) foi utilizado na base de dados, foram considerados os dados reais das dez primeiras rodadas e, em seguida, foram previstas as probabilidades para vitórias de mandante, visitantes e empates.

A figura 1 faz um comparativo entre distribuição de probabilidades de vitórias do mandante, empate e vitória do visitante, respectivamente. Tais probabilidades foram obtidas através do modelo originalmente proposto.

Revista Brasileira de Futsal e Futebol
ISSN 1984-4956 ***versão eletrônica***

Periódico do Instituto Brasileiro de Pesquisa e Ensino em Fisiologia do Exercício

w w w . i b p e f e x . c o m . b r / w w w . r b f f . c o m . b r

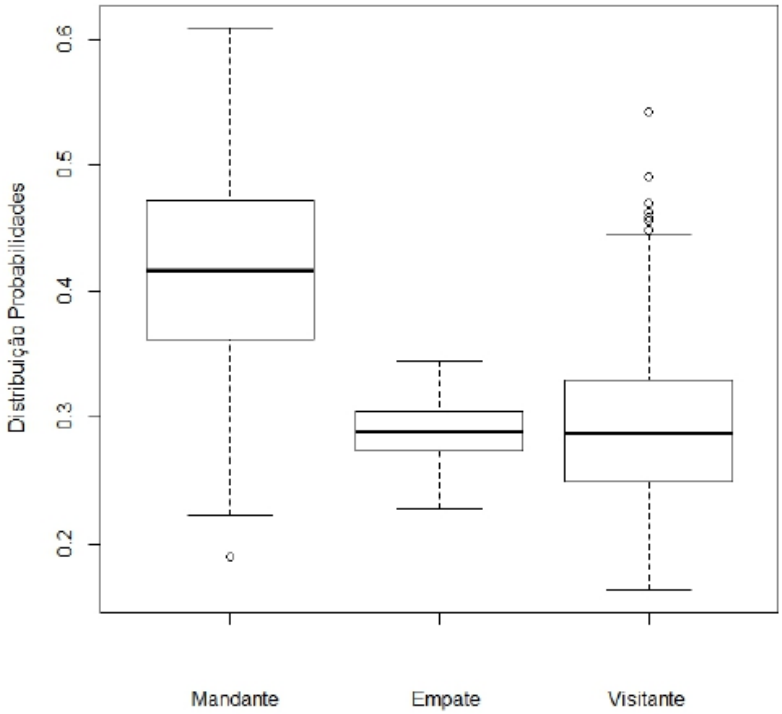

Figura 1 - Probabilidades para vitórias de mandante, visitantes e empates.

A análise gráfica mostra que o modelo desenvolvido por Martins e Duarte (2014) confirma em suas previsões um predomínio de forças para os mandantes das partidas. As probabilidades de vitória do mandante apresentam os maiores valores, porém, também apresentam grande variabilidade. Note ainda que, para pouco mais de 50% das execuções, o modelo apresenta mandantes com probabilidade de vitória superiores a 0,4. Os campeonatos disputados no Brasil são marcados por equilíbrio entre as equipes, segundo os especialistas no esporte. O modelo proposto, de certo modo, confirma este equilíbrio. Cerca de 50% das simulações resultam em probabilidades entre 0,25 e 0,32 para a vitória do visitante.

Uma análise mais específica seria a estratificação de resultados separando o resultado real e o resultado proposto através da modelagem e comparando a resposta de resultado com as probabilidades, ora de vitorias de visitantes e mandantes, ora de empates. Em outras palavras, fixado o estudo de uma das probabilidades previstas pelo modelo, por exemplo a probabilidade de vitória do mandante, três gráficos são apresentados com os três possíveis resultados verificados pelo modelo e também nos dados reais. A figura 2 ilustra detalhadamente tal análise.

Considerando os resultados reais, o equilíbrio do campeonato fica bem evidente, pois os resultados se distribuem quase de forma uniforme. Já os resultados simulados seguem uma concepção teórica clássica para o resultado do jogo. O modelo tende a prever que ocorra a vitória do time visitante quando as probabilidades de vitória do mandante são baixas. O inverso ocorre com as probabilidades de vitória do visitante com a previsão de vitória do mandante. Um ponto crítico é verificado nas situações de empate. Fica claro que o modelo não prevê altas probabilidades de empate, as simulações indicaram que apenas 14% dos jogos terminariam empatados, enquanto o percentual real foi de cerca de 25%.

Esta análise descritiva deixa clara a necessidade de utilização de algum mecanismo de correção para previsão de empates, mais que isso, o atual mecanismo em uso ainda não é perfeitamente adequado. Uma avaliação inicial, considera que seria ideal que o modelo fosse capaz de prever altas probabilidades para a vitória do mandante, probabilidades médias para os empates e baixas probabilidades para a vitória

Revista Brasileira de Futsal e Futebol
ISSN 1984-4956 ***versão eletrônica***

Periódico do Instituto Brasileiro de Pesquisa e Ensino em Fisiologia do Exercício

w w w . i b p e f e x . c o m . b r / w w w . r b f f . c o m . b r

da equipe visitante. Por outro lado, é fácil verificar que isso não ocorre.

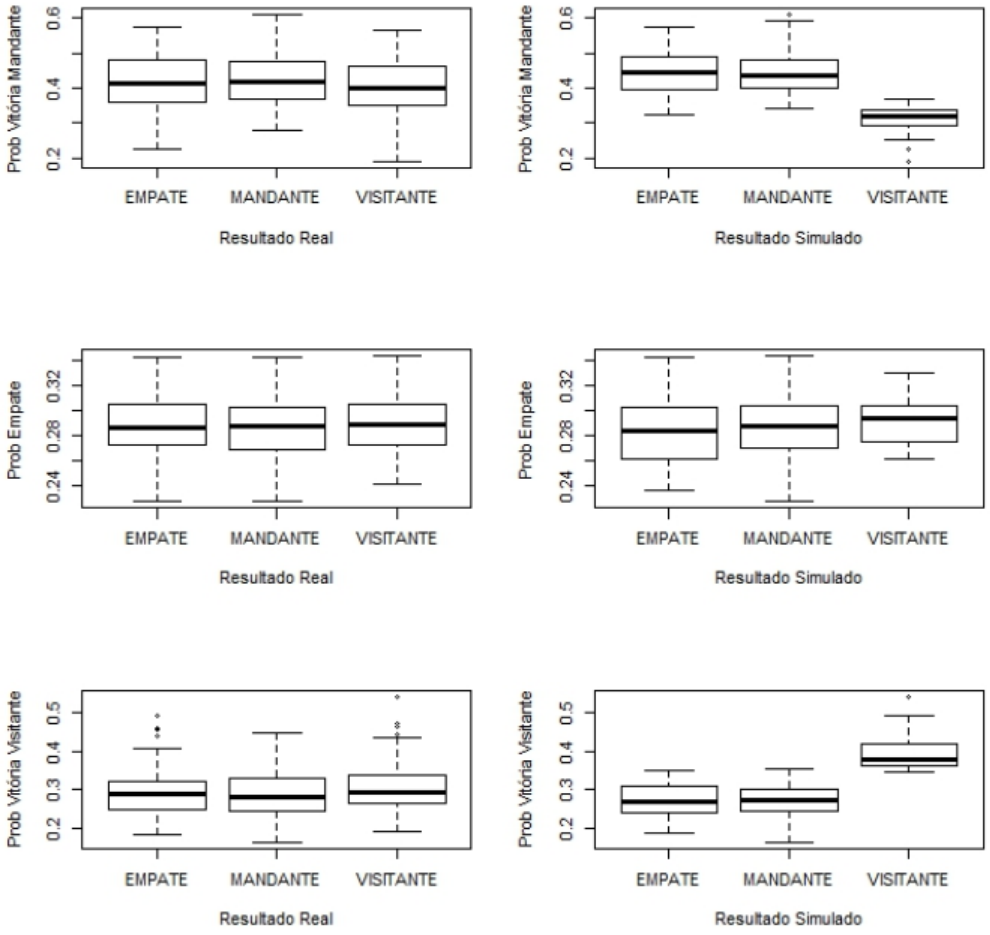

Figura 2 - Distribuição das probabilidades × resultados reais e simulados.

Outra verificação relevante foi quanto a saber se as taxas e as probabilidades apresentam concordância. Foi feito um comparativo das com as probabilidades fornecidas pelo modelo. A figura 3 traz esta comparação e como esperado, a probabilidade de vitória do mandante cresce à medida que a taxa de gols da equipe mandante cresce, e naturalmente a probabilidade de vitória do mandante cai com o aumento da taxa de gols da equipe visitante. Resultado semelhante é verificado para as taxas de gols e a probabilidade de vitória do visitante.

Um resultado bastante chamativo é a aparente falta de relação entre a taxa de gols da equipe visitante e a probabilidade de empate. Note que, independentemente do valor da taxa, a probabilidade gira em torno do valor 0,28. Já o aumento da taxa de gols da equipe mandante resulta em menores probabilidades de empate. Isto é um indicativo de que os empates reais ocorrem quando a equipe mandante apresenta baixa taxa de gols.

Com objetivo de procurar uma explicação ou mesmo um valor que se adapte à cota de correção para empates, optou-se, primeiramente, por verificar se o percentual de empates poderia ser explicado por demais informações do jogo. Foi construído um conjunto de dados com as seguintes variáveis %empates , %vitórias mandantes, gols mandantes e gols visitantes.

As variáveis utilizadas foram separadas por rodadas do campeonato e o objetivo foi tentar explicar a variável através das demais variáveis para cada rodada.

Revista Brasileira de Futsal e Futebol
ISSN 1984-4956 ***versão eletrônica***

Periódico do Instituto Brasileiro de Pesquisa e Ensino em Fisiologia do Exercício

w w w . i b p e f e x . c o m . b r / w w w . r b f f . c o m . b r

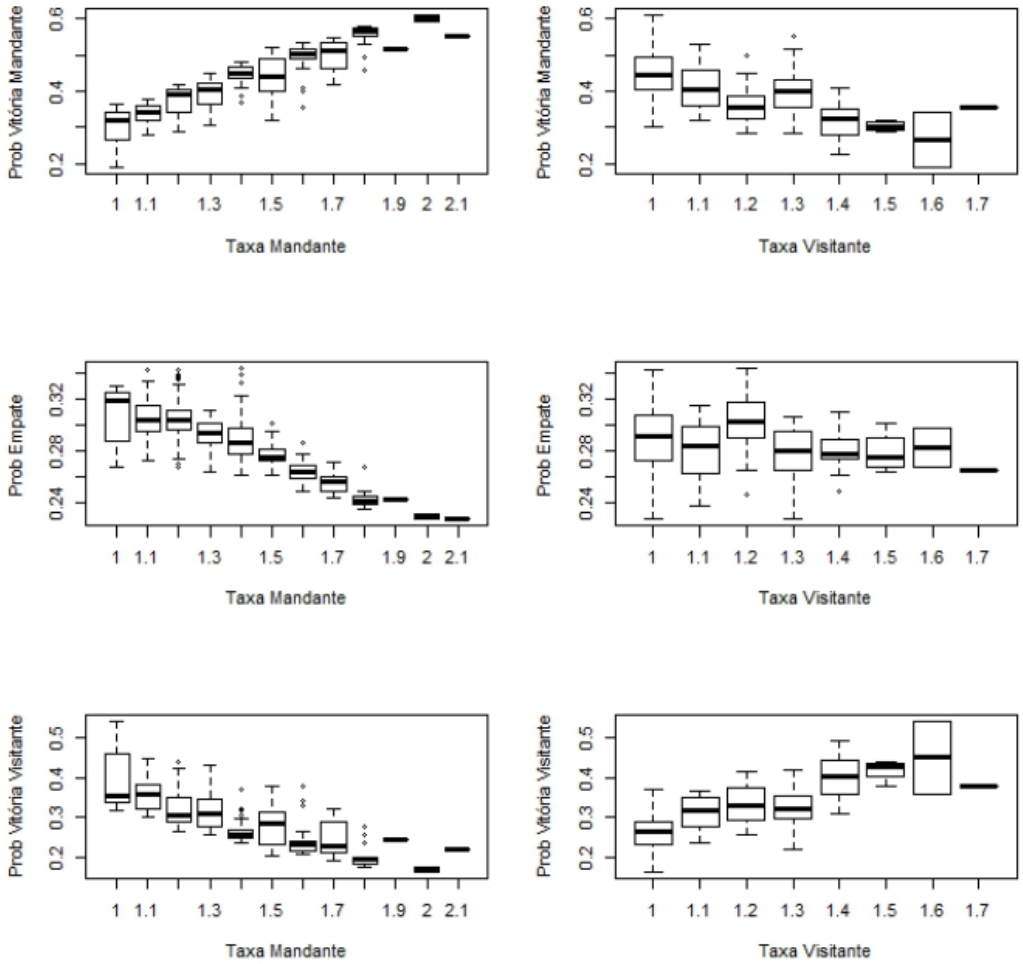

Figura 3 - Distribuição das taxas de gols × probabilidades.

A tentativa de ajuste começou pelo conjunto de dados com as dez primeiras rodadas, depois com as onze primeiras rodadas e assim sucessivamente até o modelo que utiliza todas as rodadas. Este estudo é realizado com interesse em predizer o volume de empates previstos em uma rodada qualquer do campeonato. Sua realização é feita com um campeonato já totalmente executado e o interesse é verificar a qualidade desse procedimento preditivo.

A considerar que o objetivo é predizer uma proporção, ou seja, uma medida pertencente ao intervalo (0,1), um modelo linear generalizado Beta seria uma proposta relevante para esse propósito (Cribari-Neto e Zeileis, 2010; Ferrari e Cribari-Neto, 2004;). O modelo utilizado se baseou na função de ligação logit (por uma questão de melhor adaptação), e o conjunto de avaliações subjacentes ao ajuste de tal proposição (análise de deviance, existência de pontos aberrantes ou de alavancagem) demonstrou que a escolha desse modelo seria uma possibilidade admissível.

DISCUSSÃO

De posse do ajuste do modelo linear generalizado Beta, agora será constituído o planejamento para a introdução dessa nova informação na estratégia de predição dos resultados das partidas. Inicialmente vamos relembrar que o modelo original de predição de resultados de partidas de futebol proposto por Martins e Duarte (2014) admitia que os números de gols marcados pelas equipes mandantes e visitantes constituíam um vetor aleatório bidimensional com duas coordenadas distribuídas conforme Poisson trucada.

Será denominado como Mecanismo 1, um mecanismo igual ao original, porém com

Revista Brasileira de Futsal e Futebol
ISSN 1984-4956 ***versão eletrônica***

Periódico do Instituto Brasileiro de Pesquisa e Ensino em Fisiologia do Exercício

w w w . i b p e f e x . c o m . b r / w w w . r b f f . c o m . b r

cota definida por cota = 1,07φ, em que φ é o percentual de empates ocorridos até a rodada imediatamente anterior à rodada da previsão. Já o Mecanismo 2, é um mecanismo igual ao original, a menos do percentual φ, que neste caso será substituído por $\hat{\varphi}$, uma estimativa para o percentual de empates que ocorrerá na rodada a ser prevista, com base na utilização do modelo linear generalizado Beta.

As variáveis explicativas %vitórias mandantes, gols mandantes e gols visitantes para a rodada que será prevista (e obviamente ainda não ocorreu) serão desconhecidas, portanto, o procedimento de utilização precisa ser descrito. Considere a situação de prever os resultados para a n-ésima rodada que ainda não ocorreu, no cenário que as n-1 rodadas anteriores já ocorreram. Inicialmente deve-se ajustar o modelo Beta para predizer o percentual de empates considerando os dados das n-1 rodadas conhecidas.

Para a n-ésima rodada, obtenha as distribuições Poisson truncadas para cada equipe em cada uma das partidas e execute uma série de k simulações de Monte Carlo e defina as probabilidades estimadas Vm, E e Vv. De posse dessas informações, considere que a médias das probabilidades Vm na rodada a ser prevista seja uma aproximação confiável para a variável %vitórias mandantes, na referida rodada, analogamente seja a soma das taxas das distribuições de Poisson truncadas dos mandantes como uma boa aproximação para o número de gols mandantes e a soma das taxas das distribuições de Poisson truncadas dos visitantes como uma aproximação para o número de gols visitantes. Posteriormente esses valores são aplicados ao modelo linear generalizado Beta produzindo a estimativa para o percentual de empates na rodada que se busca prever, dado por $\hat{\varphi}$, assim o mecanismo 2 terá cota definida por cota = $\hat{\varphi}$.

Os resultados mais gerais para aplicação desses dois mecanismos no dados do campeonato Brasileiro de 2015 podem ser visualizados na tabela 1, na qual a primeira coluna determina o mecanismo que está sendo utilizado, a segunda coluna apresenta a contagem geral de previsões corretas em 280 partidas, a terceira coluna apresenta a proporção de acerto de previsões nas partidas, a quarta coluna apresenta a contagem de rodadas nas quais pelo menos 40% dos resultados foram previstos corretamente. Vale lembrar que o total de rodadas de previsão é 28, visto que das 38 rodadas do campeonato, as dez rodadas iniciais foram utilizadas para calibração de parâmetros de inicialização da Poisson truncada.

Com base nas informações contidas na tabela 1, fica clara a superioridade do mecanismo 2 com respeito ao mecanismo 1. Essa informação é confirmada com base nos dados do Campeonato Brasileiro da série A de 2015.

Tabela 1 - Resultados gerais para 2015.

Mecanismo	Número de acertos	% de acertos	Rodadas com mais de 40% de acerto	% de rodadas com mais de 40% de acerto
1	87	0,310714	11	0,392857
2	99	0,353571	12	0,428571

Tabela 2 - Resultados pormenorizados para 2015.

Mecanismo	Vitórias mandantes		Empates		Vitórias visitantes	
	Número de acertos	% de acertos	Número de acertos	% de acertos	Número de acertos	% de acertos
1	35	0,246479	37	0,544118	15	0,214286
2	55	0,387324	29	0,426471	15	0,214286

Um fato novo surge ao verificar que o novo mecanismo proposto, o mecanismo 2 é superior, que por si só já representa um resultado bastante relevante. Porém, o efeito central desta constatação está no fato de que o mecanismo 1 utiliza a cota definida por cota = 1,07φ, no qual o valor 1,07 é obtido via escolha ad-hoc, sem qualquer ligação efetiva com os dados conhecidos. O mecanismo 2 não carece de tal escolha, pois ele possui uma estratégica para obtenção da cota completamente dependente dos dados.

Uma análise pormenorizada pode ser encontrada na tabela 2, na qual os dados são separados com os casos de previsões para

Revista Brasileira de Futsal e Futebol
ISSN 1984-4956 ***versão eletrônica***

Periódico do Instituto Brasileiro de Pesquisa e Ensino em Fisiologia do Exercício

w w w . i b p e f e x . c o m . b r / w w w . r b f f . c o m . b r

situações em que o resultado real foi vitória do mandante, empate e vitória do visitante.

Ao analisar a tabela 2 inicialmente verifica-se que os dois mecanismos possuem o mesmo grau de eficiência em prever resultados de vitórias de visitantes. Já na predição de empates, o mecanismo 1 é superior ao outro mecanismo.

Por fim, quando consideradas as vitórias de mandantes o modelo 1 é deficitário de forma significativa, esse é um ponto bastante negativo, visto que historicamente o volume de vitórias de mandantes é alto, ou seja, um modelo que peca em prever vitórias de mandantes é sim um modelo inadequado.

O mecanismo 2 apresenta uma melhora efetiva na predição de resultados de vitórias de mandantes.

Além disso, fica claro que a superioridade do mecanismo 1 para prever empates tem um custo muito elevado, que fica refletida de forma peremptória nas vitórias dos mandantes.

O conjunto de informações contidas nas tabelas 1 e 2 evidencia a qualidade do novo mecanismo proposto para os dados sob investigação. A fim de corroborar tal conclusão, novas análises foram realizadas considerando os dados do Campeonato Brasileiro da série A de 2013 que já haviam sido utilizados por Martins e Duarte (2014). A tabela 3 apresenta as discussões do Campeonato Brasileiro da série A de 2013 para os mecanismos 1 e 2.

A segunda análise (veja tabela 3) novamente denota uma superioridade para o mecanismo 2 em relação ao mecanismo 1. Novamente uma análise pormenorizada é apresentada na tabela 4.

O mecanismo 2, nesse caso (veja tabela 4), repetiu sua eficiência em prever corretamente os empates, mantendo praticamente números iguais aos do mecanismo 1, isso na prática é o alvo central dessa pesquisa, entretanto o mecanismo 2 apresenta resultados superiores na predição de vitórias de mandantes. Esse conjunto de constatações ilustram a superioridade do novo mecanismo proposto.

Tabela 3 - Resultados gerais para 2013.

Mecanismo	Número de acertos	% de acertos	Rodadas com mais de 40% de acerto	% de rodadas com mais de 40% de acerto
1	103	0,367857	18	0,642857
2	111	0,396429	18	0,642857

Tabela 4 - Resultados pormenorizados para 2013.

Mecanismo	Vitórias mandantes		Empates		Vitórias visitantes	
	Número de acertos	% de acertos	Número de acertos	% de acertos	Número de acertos	% de acertos
1	56	0,417910	26	0,329114	21	0,313433
2	66	0,492537	24	0,303798	21	0,313433

CONCLUSÃO

A complexidade envolvida em problemas de previsão de resultados esportivos é notória e já vastamente conhecida. Isso fica ainda mais claro quando se considera a previsão de resultados em partidas de futebol.

Os modelos discutidos nesse trabalho são de grande simplicidade e de fácil implementação, considera-se apenas e tão somente o conhecimento prévio de resultados em partidas já realizadas.

A eficiência do modelo construído através da distribuição Poisson truncada à direita já era reconhecida de antemão, bem como o procedimento outrora produzido para estimar as taxas associadas à distribuição Poisson. Esse procedimento é capaz de reduzir a superestimação para a probabilidade associada ao valor 0, um problema conhecido nas tentativas de estimar resultados de partidas esportivas através da distribuição Poisson.

Os valores e contribuições associados ao modelo propriamente proposto, mesmo considerando a reconhecida qualidade do modelo anteriormente utilizado, apresenta uma alternativa eficaz acerca da deficiência na descrição das cotas para correção de empates, que constituía um incômodo na proposição. O presente estudo revela uma

Revista Brasileira de Futsal e Futebol
ISSN 1984-4956 ***versão eletrônica***

Periódico do Instituto Brasileiro de Pesquisa e Ensino em Fisiologia do Exercício

w w w . i b p e f e x . c o m . b r / w w w . r b f f . c o m . b r

importante contribuição para sanar tal deficiência, a cota produzida através do modelo linear generalizado Beta.

Os dois mecanismos apresentaram resultados condizentes com os volumes de vitórias de mandantes, visitantes e empates nos Campeonatos Brasileiros da série A de 2013 e 2015. A qualidade dos resultados da nova proposta é claramente decorrente da melhoria no formato de proposição da cota para correção de empates imposta ao estimador de resultados de partidas.

Apesar da cota simplesmente associada ao volume de empates já ocorridos no campeonato já se adaptar bem, é possível verificar o efeito de melhoria da cota corrigindo resultados que estimadores mais ingênuos não seriam capazes de fazê-lo.

A possibilidade de previsão de resultados de campeonatos completos de futebol continua sendo um foco a ser alcançado e se torna cada vez mais plausível de ser executado de forma eficaz se considerados os mecanismos propostos aqui. Outras propostas de continuidade de estudo podem ser observadas na possibilidade de predição de placares de partidas e não somente prever vencedores e empates. Esse é claramente um objetivo extremamente mais sofisticado, devido à magnitude de seu espaço de possíveis resultados.

REFERÊNCIAS

1-Alves, A. M. e colaboradores. Logit models for the probability of winning football games. Pesquisa Operacional. Vol. 31. p. 459-465. 2011.

2-Brillinger, D. An analysis of chinese super league partial results. Science in China Series A: Mathematics. Vol. 52. p. 1139-1156. 2009.

3-Cribari-Neto, F.; Zeileis, A. Beta regression in R. Journal of Statistical Software. Articles. Vol. 34. Num. 2. p. 1-24. 2010.

4-Faria, F. F. Análise e Previsão de Resultados de Partidas de Futebol. Universidade Federal do Rio de Janeiro. Tese de Doutorado. 2005.

5-Ferrari, S. L. P.; Cribari-Neto, F. Beta regression for modelling rates and proportions. Journal of Applied Statistics. Vol. 31. Num. y. p. 799-815. 2004.

6-Karlis, D.; Ntzoufras, I. On model soccer data. Student. Vol. 3. Num. 4. p. 229-244. 2000.

7-Karlis, D.; Ntzoufras, I. Analysis of sports data by using bivariate poisson models. Journal of the Royal Statistical Society. Series D (The Statistician). Vol. 52. Num. 3. p. 381-393. 2003.

8-Knorr-Held, L. Dynamic rating of sports teams. Journal of the Royal Statistical Society. Series D (The Statistician). Vol. 49. Num. 2. p. 261-276. 2000.

9-Liu, F.; Zhang, Z. Predicting soccer league games using multinomial logistic models. Course Project. 2008.

10-Macedo, P. A. P.; Silva, C. D. Prediction results in the brazilian championship 2012 series a. The Brazilian Journal of Soccer Science. Vol. 7. Num. 2. p. 35-41. 2014.

11-Martins, H. S. R.; Duarte, A. R. Modelagem dinâmica de partidas de futebol. Revista da Estatística da Universidade Federal de Ouro Preto. Vol. 3. Num. 2. p. 157-169. 2014.

12-Rotshtein, A. P.; Posner, M.; Rakityanskaya, A. B. Football predictions based on a fuzzy model with genetic and neural tuning. Cybernetics and Systems Analysis. Vol. 41. Num. 4. p. 619-630. 2005.

13-Rue, H.; Salvesen, Ø. Prediction and retrospective analysis of soccer matches in a league. Journal of the Royal Statistical Society. Series D (The Statistician). Vol. 49. Num. 3. p. 399-418. 2000.

14-Souza Jr., O. G.; Gamerman, D. Previsão de partidas de futebol usando modelos dinâmicos. In: Anais do XXXVI evento da Sociedade Brasileira de Pesquisa Operacional. SBPO. 2004. p. 650-659.

Recebido para publicação em 26/07/2018
Aceito em 06/01/2019