Revista da Estatística UFOP, Vol III(2), 2014, ISSN 2237-8111
Edição especial: 59ª Reunião Anual da Regional Brasileira da Sociedade Internacional de Biometria - RBRAS

MODELAGEM DINÂMICA DE PARTIDAS DE FUTEBOL

Helgem de Souza Ribeiro Martins[1], **Anderson Ribeiro Duarte**[1]

Resumo: *O futebol se enquadra entre os esportes de mais difícil previsibilidade, ou seja, no qual a ocorrência de resultados em que equipes inferiores superam as equipes melhores se faz mais comum. Nesse trabalho busca-se obter previsibilidade para o resultado de partidas de futebol através de um modelo Poisson truncado à direita e uma cota dinâmica para a determinação de empates. Um modelo de simulação é apresentado e resultados encorajadores são discutidos para o Campeonato Brasileiro de Futebol da Série A e também para a Premier League inglesa.*

Palavras-chave: *Futebol, Previsibilidade, Poisson truncada, Campeonato Brasileiro de Futebol, Premier League inglesa.*

Abstract: *The football is included among the hardest sports to predictability. In this work we seek to achieve predictability to outcome of matches through a truncated Poisson and a dynamic threshold to determination of tie. A simulation model and results are discussed for the Brazilian Championship and the English Premier League.*

Keywords: *Football, Predictability, Truncated Poisson, Brazilian Championship, English Premier League.*

1 Introdução

O futebol é, notoriamente, uma paixão mundial. É surpreendente a emoção envolvida, não apenas entre os praticantes, mas principalmente entre os aficionados que o acompanham. Trata-se de um assunto com vasto espaço nas mídias escritas e faladas. Até mesmo aqueles que não são fãs de primeira ordem, em algum momento, acabam demonstrando algum envolvimento. Em diversos países, o esporte gera vultuosas movimentações financeiras em casas de apostas. Em alguns desses países, as apostas são legalizadas, em outros não. Busca-se o acerto do resultado de uma dada partida, seu placar exato e até mesmo o acerto daquele atleta que abrirá o placar, o instante de tempo que ocorrerá o primeiro gol, dentre outras diversas possibilidades de apostas.

Dado o grande volume de recursos financeiros envolvido nas equipes de primeiro escalão, seria plenamente esperado, por alguns, que um modelo simples que associasse a chance de vitória ao volume de receitas ou então ao volume financeiro investido se tornaria bastante eficaz. Entretanto, os resultados muitas vezes contradizem isto de forma contundente. Em muitas oportunidades, é possível observar, equipes de grande investimento serem derrotadas por equipes inexpressivas do ponto de vista econômico.

O futebol se enquadra entre os esportes de mais difícil previsibilidade, ou seja, no qual a ocorrência de resultados em que equipes inferiores superam as equipes melhores se faz mais comum. Trata-se de uma situação mais típica dos esportes coletivos e de pontuação baixa. Em

[1]DEEST - UFOP.

outras palavras, é mais fácil observar um resultado atípico (*zebra*) no futebol que, por exemplo, no basquete.

Estes fatores são os que mais despertam interesse nesta área, tornando extremamente difícil produzir modelos e estratégias na busca da previsibilidade. Os esportes, em geral, e não somente o futebol, são cada vez mais alvo de especialistas que buscam em modelos matemáticos e estatísticos os caminhos para explicar e prever os resultados. Trabalhos vêm sendo desenvolvidos, não somente com o futebol, mas com outros esportes também, como por exemplo, [5, 7, 8].

Com o grande avanço das técnicas estatísticas e também da capacidade de processamento de dados dos computadores nos últimos anos, cada vez mais se torna possível tentar tratar dados dessa natureza. A produção de modelos eficientes para previsibilidade em jogos de futebol não somente serve para estimar o resultado de uma partida que ainda ocorrerá, mas também para estimar as chances de obtenção de títulos ou então de rebaixamento entre divisões para as equipes participantes ao final de cada campeonato. O controle sobre a natureza desses resultados notoriamente envolve cifras inimagináveis.

Os modelos clássicos, em geral, utilizam séries históricas com janela de tempo curta para tentar incluir estes efeitos na análise para os resultados das partidas futuras. Esta é a principal justificativa para se buscar uma estratégia de modelagem dinâmica tentando captar as alternâncias das equipes ao decorrer da disputa do campeonato. Nesse trabalho busca-se obter previsibilidade somente sobre o resultado de partidas, deve ficar claro que uma extensão óbvia é obter previsibilidade sobre o resultado de campeonatos, mas se enquadra em objetivos futuros de pesquisa.

A previsão em campeonatos completos demanda uma metodologia superior à necessária para a obtenção de resultados de partidas. Deve-se notar que ao buscar previsões em resultados de campeonatos, diversas partidas serão simuladas e em algum momento, o intervalo de tempo entre os dados reais (partidas já realizadas) e as partidas que serão simuladas se torna suficientemente grande, dificultando em demasia a capacidade de previsão dos resultados. O principal foco deste trabalho está em obter previsões de resultados apenas de partidas, que serão analisados de forma agregada em: vitórias, empates e derrotas.

O trabalho se encontra organizado da seguinte forma: a seção 2 apresenta uma revisão bibliográfica de métodos apresentados nessa área de estudo, mostra discussões sobre a distribuição Poisson truncada à direita e a técnica para a estimação dos parâmetros associados às distribuições Poisson para o problema em estudo, além de discutir e detalhar o modelo de simulação e as estratégias envolvidas; a seção 3 relata o conjunto de experimentos que será realizado, contemplando o Campeonato Brasileiro da série A e também a *Premier League* inglesa (primeira divisão do campeonato inglês de futebol); conclusões e também propostas de continuidade para esse estudo são relatadas na seção 4.

2 Material e métodos

Diversas modelagens são propostas com o intuito de prever o resultado de partidas de futebol. Em Fang & Zheng [6] é apresentada uma abordagem via modelo de regressão multinomial, enquanto em Brillinger [1] um modelo trinomial baseado nos três possíveis resultados de uma partida de futebol é utilizado. Em alguns trabalhos desta temática as análises são realizadas com a utilização de modelos Bayesianos, como em Farias [2] e Souza & Gamerman [8], porém, a modelagem mais utilizada neste contexto são as baseadas em modelar o volume de gols através de uma distribuição de Poisson e também suas variações.

Em Karlis & Ntzoufras [4] são verificados os pressupostos de independência entre os gols anotados pelos adversários em uma partida, assumindo que a distribuição conjunta dos gols de cada partida segue uma distribuição bivariada de Poisson, independente do parâmetro de correlação e avalia a viabilidade da utilização da referida distribuição em modelos discretos para simulação de resultados de partidas esportivas. Em Karlis & Ntzoufras [3], tal resultado é novamente explorado e aplicado na modelagem de 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.

O modelo proposto neste trabalho utiliza uma distribuição de Poisson truncada à direita, de modo à limitar o número máximo de gols em cada partida, propõe um fator de ajuste dinâmico das taxas de gols e ainda um efeito também dinâmico de correção para a previsão de resultados de empate. Isso se deve o fato de que o futebol é um esporte cujos resultados são visivelmente influenciados pelo desempenho momentâneo, fatores extra-campo, contusões, suspensões, punições e diversos fatores que tornam o mesmo, um dos esportes mais imprevisíveis praticados pelo homem. Tal dinâmica busca captar estas oscilações e colocá-las à serviço do modelo de previsão de resultados de partidas de futebol.

2.1 Distribuição de Poisson Truncada à direita

Ao se verificar o histórico de placares em torneios de futebol ao redor do globo, percebe-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, propõe-se fixar uma cota superior à distribuição dos gols marcados por cada time. Tomando-se tal cota, 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 x_{max} a cota superior máxima de gols que tal equipe pode fazer na referida partida. As probilidades para a variável aleatória X podem ser obtidas por:

$$P(X = x|\ X \leq x_{max}) = \frac{\frac{e^{-\lambda}\lambda^x}{x!}}{\sum_{k=0}^{x_{max}} \frac{e^{-\lambda}\lambda^k}{k!}} = \frac{\lambda^x}{x! \sum_{k=0}^{x_{max}} \frac{\lambda^k}{k!}}; \quad x, k = 0, 1, 2, \cdots, x_{max} \tag{1}$$

em que λ representa a taxa da distribuição de gols da referida equipe. A taxa e a escolha da cota superior utilizada serão definidas a seguir.

2.2 Estimação de Parâmetros

Uma partida de futebol apresenta diversos parâmetros que poderiam ser analisados e incluídos em um modelo de previsão de seu resultado, porém, por parcimônia, é adequado propor a busca por uma solução mais simples. Podemos considerar de forma simplificada, que o placar de uma partida é resultado de basicamente dois fatores associados a cada uma das equipes envolvidas: a capacidade defensiva, que se reflete no número de gols sofridos, e a capacidade ofensiva, que influencia na quantidade de gols marcados. Busca-se então estimar tais fatores a partir apenas dos gols pró e contra das equipes envolvidas. Desta forma, trabalha-se apenas com a pontuação básica do futebol, o *gol*.

Diversos modelos, entre esses, todos os citados anteriormente, incluem um fator *Mando de campo*, onde é concedida à equipe mandante alguma forma de vantagem, em geral a partir de uma variável *dummy*. O modelo que está sendo proposto não incorpora este tipo de vantagem. Sendo assim, uma das estratégias que podem ser utilizadas na verificação do ajuste do modelo é o desempenho dos mandantes das partidas, dado que espera-se que os mesmos possuam probabilidade de vitória superior ao visitante na maioria dos casos.

Serão definidos os fatores para cada equipe como *Fator Ataque* e *Fator Defesa*, nos quais são representadas as capacidades ofensiva e defensiva respectivamente, da seguinte forma:

- Fator ataque: μ_n^+(Mandante) e γ_n^+(Visitante)
- Fator defesa: μ_n^-(Mandante) e γ_n^-(Visitante)

em que:
μ_n^+ = Média dos gols feitos pelo time mandante nas últimas n rodadas como mandante.
μ_n^- = Média dos gols sofridos pelo time mandante nas últimas n rodadas como mandante.
γ_n^+ = Média dos gols feitos pelo time visitante nas últimas n rodadas como visitante.
γ_n^- = Média dos gols sofridos pelo time visitante nas últimas n rodadas como visitante.

De posse de tais fatores, as taxas de gols para mandantes λ_M^+ e visitantes λ_V^+ serão definidas a partir das seguintes equações:

$$\lambda_M^+ = \max\left(1, \frac{\mu_n^+ + \gamma_n^-}{2}\right) \tag{2}$$

$$\lambda_V^+ = \max\left(1, \frac{\gamma_n^+ + \mu_n^-}{2}\right) \tag{3}$$

A taxa de gols para a equipe mandante (λ_M^+) será obtida considerando o máximo entre 1 e a média entre o fator ataque da equipe mandante e o fator defesa da equipe visitante. Analogamente a taxa de gols para a equipe visitante (λ_V^+) será obtida considerando o máximo entre 1 e a média entre o fator ataque da equipe visitante e o fator defesa da equipe mandante. A escolha pelo máximo entre as referidas médias e o valor 1 possui uma justificativa técnica. Para taxas inferiores a 1 a probabilidade associada ao valor 0 na Poisson truncada à direita se torna alta o suficiente para previlegiar muito este resultado, este fato não é adequado para simulação de resultados de partidas de futebol. Em outras palavras, está se afirmando que o valor 0 é obviamente um resultado possível, porém não tão provável quanto se tornaria na estimação dos valores λ sem a cota inferior 1. Este efeito majorante na probabilidade associada ao valor 0 pode ser claramente visualizado na Figura 1.

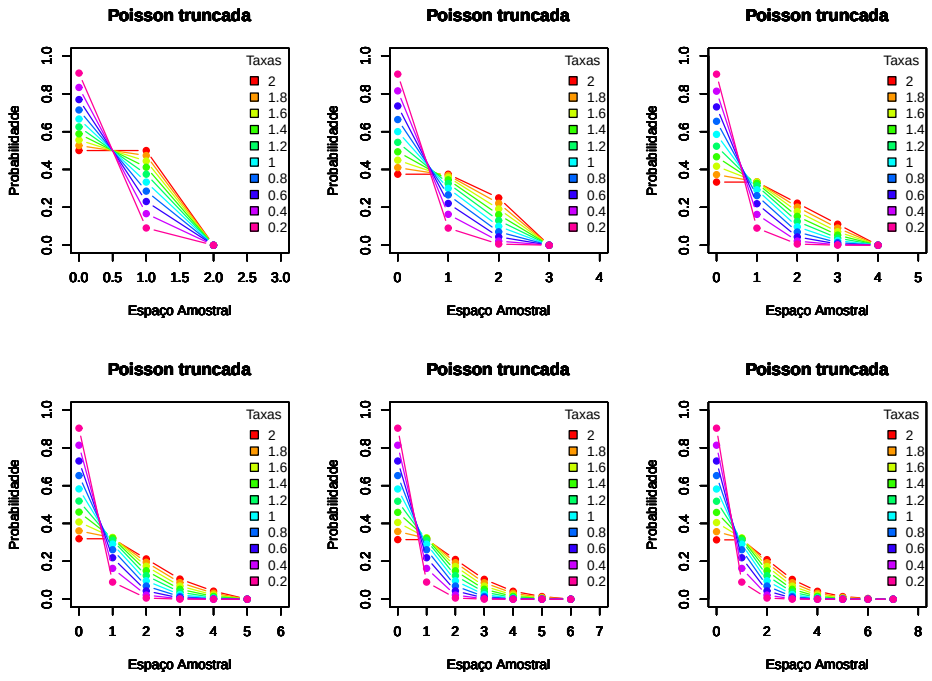

Figura 1: Distribuições de probabilidade truncadas à direita.

O número de rodadas (n) consideradas da série histórica deve ser capaz de captar informações efetivas sobre a equipe, porém não deve utilizar informações de um passado muito distante. O

intuito é o de observar a capacidade da equipe na atualidade, ou seja, para valores elevados de (n) seriam utilizadas informações de uma equipe potencialmente diferente da atual, isto por causa da alta rotatividade dos atletas na equipe-base que disputa as partidas. Nesse o trabalho, o valor (n) será fixado em 5, pois a intenção do modelo é considerar intervalos 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. Desta forma, define-se a "fase" de cada time como a janela de tempo com espaço de dois meses.

2.3 Definição do Modelo

O modelo de simulação computacional para o resultado de partidas de futebol será construído utilizando-se a distribuição acumulada da Poisson truncada à direita, definindo-se o tamanho da série histórica considerada e o número no qual se truncará a distribuição de referência. Conforme mencionado, serão consideradas as últimas 5 rodadas como mandante para a equipe que jogará em seus domínios, e as últimas 5 rodadas como visitante para o adversário que joga fora de sua sede.

Além das taxas definidas anteriormente, o modelo depende da escolha do ponto de truncamento para a distribuição Poisson. Para tanto, seja $\max G_M$ o número máximo de gols que a equipe mandante fez em uma única partida como mandante dentre as partidas consideradas para a obtenção de λ_M^+. Seja ainda, $\max G_V$ o número máximo de gols que a equipe visitante fez em uma única partida como visitante dentre as partidas consideradas para a obtenção de λ_V^+. O valor de truncamento será definido por $x_{max} = \max(\max G_M, \max G_V) + 1$, neste caso, está sendo definido um único ponto de truncamento para as duas equipes envolvidas na partida, nada impede a possibilidade de pontos de truncamento distintos entre as equipes. Teste para valores distintos foram executados revelando que o impacto é bastante pequeno. Uma dúvida pode pairar sobre a escolha do máximo entre $\max G_M$ e $\max G_V$ e sobre a adicionar uma unidade para a obtenção de x_{max}. As duas escolhas se devem ao fato de não limitar as equipes a somente realizar, em uma partida simulada, um número de gols que ela já tenha feito em uma partida do passado. Busca-se admitir portanto, que uma equipe pode ser, até mesmo, um pouco superior aos dados de sua série histórica.

2.4 Modelo de Simulação

O modelo de simulação para partidas de futebol que está sendo proposto será executado em duas etapas, uma primeira etapa de construção das distribuições e uma segunda etapa de efetivo processo de simulação utilizando as distribuições construídas na etapa anterior. No procedimento de construção das distribuições, os fatores $\mu_5^+, \mu_5^-, \gamma_5^+, \gamma_5^-$ são obtidos para as duas equipes envolvidas na partida. De posse desses valores, são obtidas as taxas λ_M^+ e λ_V^+ e o ponto de truncamento x_{max}, chegando assim a uma distribuição bivariada ($\texttt{gols}_\texttt{M}$,$\texttt{gols}_\texttt{V}$), note que está sendo considerado que $\texttt{gols}_\texttt{M}$ e $\texttt{gols}_\texttt{V}$ são duas variáveis aleatórias Poisson truncada à direita em x_{max} com parâmetros λ_M^+ e λ_V^+ respectivamente, e independentes.

A segunda etapa do procedimento consiste em reproduzir por meio de simulação computacional uma realização do vetor aleatório ($\texttt{gols}_\texttt{M}$,$\texttt{gols}_\texttt{V}$). Ao final de k repetições da segunda etapa do procedimento, as contagens dos casos em que $\texttt{gols}_\texttt{M} > \texttt{gols}_\texttt{V}$, $\texttt{gols}_\texttt{V} > \texttt{gols}_\texttt{M}$ e $\texttt{gols}_\texttt{M} = \texttt{gols}_\texttt{V}$ são armazenadas para serem tratadas como possíveis estimativas das chance de vitórias de mandantes, visitantes ou empates para cada partida simulada.

Um pseudo-código contendo o modelo simplificado do algoritmo computacional de simulação é apresentado no Algoritmo 1. O algoritmo proposto foi desenvolvido na linguagem de programação $C++$, e todos os testes foram executados em um computador Intel Core I5 2410M, 2.3GHz com 4GB de memória RAM.

Algoritmo 1 Simulador de Partidas de Futebol - SPF

1: Leitura dos dados;
2: Cálculo das taxas de gols λ_M^+ e λ_V^+;
3: Cálculo dos pontos de truncamento T;
4: Construção das Distribuições de gols ($\texttt{gols}_\texttt{M}$,$\texttt{gols}_\texttt{V}$);
5: **para** $\texttt{n}_\texttt{sim} = 1$ até k **faça**
6: Simula uma realização de $\texttt{gols}_\texttt{M}$;
7: Simula uma realização de $\texttt{gols}_\texttt{V}$;
8: **se** $\texttt{gols}_\texttt{M} > \texttt{gols}_\texttt{V}$ **então** $\texttt{V}_\texttt{M} = \texttt{V}_\texttt{M} + 1$
9: **se** $\texttt{gols}_\texttt{M} < \texttt{gols}_\texttt{V}$ **então** $\texttt{V}_\texttt{V} = \texttt{V}_\texttt{V} + 1$
10: **se** $\texttt{gols}_\texttt{M} = \texttt{gols}_\texttt{V}$ **então** $\texttt{E} = \texttt{E} + 1$
11: **fim para**
12: **imprime**($\texttt{V}_\texttt{M}$, $\texttt{E}$, $\texttt{V}_\texttt{V}$)

2.5 Estimador de resultados de partidas

De posse do modelo de simulação de partidas, resta obter o principal interesse do trabalho, a estimativa para o resultado de uma partida cujo placar final é desconhecido. Mesmo considerando toda a base metodologica que já foi discutida, ainda não foi esclarecido como se dará tal procedimento. Algumas situações podem ser aventadas, para tanto esse processo será iniciado de forma bastante superficial.

Defina $\theta = (x_1, x_2, x_3)$ um vetor de coordenadas binárias cuja soma é sempre 1, o objetivo é representar o resultado da partida através deste vetor:

$$\theta = \begin{cases} (1,0,0) & \text{se a equipe mandante é vencedora;} \\ (0,1,0) & \text{se ocorre um empate;} \\ (0,0,1) & \text{se a equipe visitante é vencedora.} \end{cases} \tag{4}$$

Uma primeira proposta surge de forma bem natural, suponha apresentar para um fã do futebol, uma lista de partidas e a lista de taxas λ_M^+ e λ_V^+ para as referidas partidas, um estimador surgiria considerando vencedora a equipe de maior taxa e empate para os casos com taxas iguais. Esta proposta inicial leva ao estimador $\hat{\theta}_I$ dado por:

$$\hat{\theta}_I = \begin{cases} (1,0,0) & \text{se } \lambda_M^+ > \lambda_V^+; \\ (0,1,0) & \text{se } \lambda_M^+ = \lambda_V^+; \\ (0,0,1) & \text{se } \lambda_M^+ < \lambda_V^+. \end{cases} \tag{5}$$

Uma segunda opção, já requer um pouco mais de informação. Suponha possuir as taxas λ_M^+ e λ_V^+ e a proposta de utilização das distribuições Poisson truncadas à direita para a produção de um simulador de partidas. Considerando uma sequência de execuções de simulações, uma contagem simples de ocorrências vitórias de mandante (V_M), empates (E) e vitórias de visitante (V_V) seria suficiente para a produção de um novo estimador $\hat{\theta}_{II}$ dado por:

$$\hat{\theta}_{II} = \begin{cases} (1,0,0) & \text{se } \max(V_M, E, V_V) = V_M; \\ (0,1,0) & \text{se } \max(V_M, E, V_V) = E; \\ (0,0,1) & \text{se } \max(V_M, E, V_V) = V_V. \end{cases} \tag{6}$$

Uma terceira opção parte novamente dos resultados do simulador de partidas, ou seja, considera o conhecimento das taxas λ_M^+ e λ_V^+ e da proposta de utilização das distribuições Poisson truncadas à direita. Considerando uma sequência de execuções de simulações, o resultado é uma amostra de pares ordenados, com uma coordenada sendo o número de gols da equipe mandante ($\texttt{gols}_\texttt{M}$) e a outra sendo o número de gols da equipe visitante ($\texttt{gols}_\texttt{V}$) para cada execução da simulação. Um novo estimador $\hat{\theta}_{III}$ pode ser obtido considerando o par ordenado mais frequente na amostra, ou seja, a moda da amostra de pares ordenados.

$$\hat{\theta}_{III} = \begin{cases} (1,0,0) & \text{se } MO_M > MO_V; \\ (0,1,0) & \text{se } MO_M = MO_V; \\ (0,0,1) & \text{se } MO_M < MO_V; \end{cases} \tag{7}$$

em que (MO_M, MO_V) é o par ordenado mais frequente na amostra simulada.

Finalmente será apresentada a proposta de um estimador que busca avaliar todos estes valores e raciocínios, e produzir um formato que seja mais eficaz. Novamente, considerando uma sequência de execuções de simulações utilizando a Poisson truncada à direita, e a contagem simples de ocorrências vitórias de mandante (V_M), empates (E) e vitórias de visitante (V_V), o estimador proposto na equação (6) simplesmente verifica o máximo entre V_M, E, V_V, mas algumas situações ambíguas podem ocorrer. O novo estimador considera que a proposta de $\hat{\theta}_{II}$ parece adequada para estimar as vitórias de visitantes, afinal, dada a natureza do futebol, detectar que $\max(V_M, E, V_V) = V_V$ parece um efeito suficiente para corroborar as vitórias de visitantes. Já $\max(V_M, E, V_V) = E$ parece adequado para estimar os empates, porém insuficiente. Dado que qualquer modelo de previsão de partidas de futebol tende de alguma forma a aumentar a chance de vitória da equipe mandante, não é claro que $\max(V_M, E, V_V) = V_M$, obtido através do modelo, seja suficiente para confirmar a referida vitória. Em muitos casos, no cenário do estimador $\hat{\theta}_{II}$ tem-se que proporcionalmente o volume de empates nos dados de simulação é baixo. Os casos em que $\max(V_M, E, V_V) = V_M$ e a proporção $\frac{E}{(V_M+E+V_V)}$ seja suficientemente elevada serão considerados empates.

$$\hat{\theta} = \begin{cases} (1,0,0) & \text{se } \max(V_M, E, V_V) = V_M \text{ e } \frac{E}{(V_M+E+V_V)} < cota; \\ (0,1,0) & \text{se } \max(V_M, E, V_V) = V_M \text{ e } \frac{E}{(V_M+E+V_V)} \geq cota \\ & \text{ou se } \max(V_M, E, V_V) = E; \\ (0,0,1) & \text{se } \max(V_M, E, V_V) = V_V. \end{cases} \tag{8}$$

Obviamente, estabelecer uma cota para adequada para as percentagens $\frac{E}{(V_M+E+V_V)}$ na qual as situações de $\max(V_M, E, V_V) = V_M$ seriam transformadas em resultados de empate não parece um problema trivial. O primeiro teste foi convencionar que essa cota seria estabelecida pelo percentual de empates (aqui definido por φ) ocorridos nos dados reais anteriores à rodada que está sendo simulada. Este teste leva a um estimador que privilegia um volume alto de empates, logo a cota deveria ser superior. Diversos testes empíricos foram realizados e a cota melhor sucedida foi $cota = 1{,}07\varphi$. Mais detalhes sobre a utilização desta cota serão apresentados na seção seguinte.

3 Resultados e discussão

O modelo descrito anteriormente foi aplicado aos jogos do Campeonato Brasileiro de Futebol da Séria A de 2013 e também para partidas da Premier League (primeira divisão inglesa) 2013/14. É importante ressaltar que o objetivo, presente neste estudo, não é o de simular campeonatos, mas sim simular partidas em campeonatos. A simulação de campeonatos requer, a partir de algum instante da simulaçao, que o distância temporal entre os dados de entrada para a simulação e as partidas que serão simuladas se tornem muito longos. Não estamos afirmando sobre inviabilidade desse procedimento, mas algumas técnicas ainda precisam ser constituídas. O procedimento de simulação de partidas é mais simples e será executado considerando dados reais ocorridos antes da partida que se busca simular.

Para o Campeonato Brasileiro da Séria A de 2013, vale resaltar que todas as suas partidas já foram concluídas, ou seja, o procedimento será de simular uma rodada de partidas e comparar o com os resultados reais para a respectiva rodada. Já a Premier League 2013/14 ainda se encontra em andamento, mas algumas de suas rodadas já realizadas serão simuladas através do mesmo mecanismo usado no Campeonato Brasileiro. Os dados associados aos resultados

reais para o Campeonato Brasileiro 2013 foram obtidos em [10], enquanto os dados referentes à Premier League 2013/14 foram obtidos em [9]

3.1 Campeonato Brasileiro 2013

A simulação do Campeonato Brasileiro 2013 se torna interessante para verificação do ajuste do modelo, pois este já se findou, possibilitando a imediata comparação entre o resultado simulado e o resultado real. O histórico deste torneio é repleto de surpresas, sendo que em diversas ocasiões, equipes apontadas como favoritas em seu início travaram batalhas contra o rebaixamento para a divisão inferior, e equipes desacreditadas, de baixo investimento financeiro, surpreenderam e brigaram pelas primeiras posições. A liga brasileira de futebol de 2013 é um exemplo perfeito do equilíbrio e imprevisibilidade de seus resultados. O campeão do ano anterior (2012), apontado por diversos especialistas no esporte, como um dos três favoritos ao título de 2013, figurou durante toda a competição entre os últimos colocados, enquanto o vencedor do certame foi uma equipe que não dispunha de grandes investimentos e prestígio no ano, mas acabou conquistando o campeonato.

Conforme já mencionado, as simulações foram realizadas apenas para partidas já realizadas. Foram simulados os resultados de jogos entre as rodadas 28 e 38. Para cada partida foram realizadas 20.000 simulações e o algoritmo é rápido o suficiente para simular os 2.200.000 jogos da rotina em tempo inferior a 20 segundos.

A Tabela 1 compara o volume de vitórias dos mandantes, empates e vitórias dos visitantes em dados simulados com os valores reais. Trata-se de uma análise para a verificação da adequabilidade do modelo proposto com a realidade do campeonato. Pode-se observar que nem todos os estimadores apresentam resultados simulados que se aproximam dos valores reais. Apenas o estimador $\hat{\theta}_{III}$ não privilegiou o resultado mais esperado, a vitória do mandante.

Tabela 1: Comparação entre a ocorrências de resultados reais e simulados (volume total).

	Vitórias de Mandantes	Empates	Vitórias de Visitantes
Ocorrências Reais	61	30	19
Ocorrências Simuladas ($\hat{\theta}_I$)	88	9	13
Ocorrências Simuladas ($\hat{\theta}_{II}$)	91	1	18
Ocorrências Simuladas ($\hat{\theta}_{III}$)	30	62	18
Ocorrências Simuladas ($\hat{\theta}$)	71	21	18

A Tabela 1 mostra que os estimadores $\hat{\theta}_I$ e $\hat{\theta}_{II}$ possuem uma tendenciosidade contra o resultado de empate. Já o estimador $\hat{\theta}_{III}$ possui uma tendenciosidade à favor do empate. O estimador $\hat{\theta}$ apresenta resultados um pouco mais condizentes em relação a distribuição dos resultados entre vitórias de mandantes, visitantes e empates. Este resultado é importante, visando produzir futuramente estimativas para o campeonato completo. Considerando que a equipe vencedora ganha três pontos e em empates cada uma das duas equipes ganha um ponto, o volume total de pontos distribuídos no campeonato, quando usando $\hat{\theta}_I$ ou $\hat{\theta}_{II}$ tende a ser superior ao total real e quando usando $\hat{\theta}_{III}$ tende a ser inferior ao total real. Em relação aos resultados de cada partida, a Tabela 2 apresenta um panorama geral.

Os estimadores $\hat{\theta}_I$ e $\hat{\theta}_{II}$ apresentaram 48% e 49% de resultados simulados iguais aos resultados reais, respectivamente. O estimador $\hat{\theta}$ apresentou exatamente 50% dos resultados simulados sendo condizentes com o resultado real da partida. O estimador $\hat{\theta}_{III}$ se mostrou inferior com uma faixa de acerto na casa de 30%. A Tabela 2 apresenta para cada estimador o percentual de acerto dentro das tentativas de estimação, ou seja, dado que um estimador considerou a quantidade V_M de vitórias de equipes mandantes e uma quantidade AV_M destas vitórias realmente se concretizaram, então o percentual é dado por $\frac{AV_M \times 100}{V_M}$. Além disso, são apresentadas as percentagens de acerto em relação ao resultado real, ou seja, para uma quantidade AV_M de acertos

Tabela 2: Análise do número de resultados simulados corretamente.

Acerto dentro das tentativas			
Estimador	Vitórias de Mandantes	Empates	Vitórias de Visitantes
$\hat{\theta}_I$	54,5%	11,1%	30,7%
$\hat{\theta}_{II}$	54,9%	0,0%	22,2%
$\hat{\theta}_{III}$	50,0%	29,0%	5,5%
$\hat{\theta}$	59,2%	42,9%	22,2%
Acerto geral			
Estimador	Vitórias de Mandantes	Empates	Vitórias de Visitantes
$\hat{\theta}_I$	78,7%	3,3%	21,1%
$\hat{\theta}_{II}$	81,9%	0,0%	21,1%
$\hat{\theta}_{III}$	24,6%	60,0%	5,3%
$\hat{\theta}$	68,9%	30,0%	21,1%

do estimador, dentre as vitórias de equipes mandantes e uma quantidade TV_M sendo o efetivo número de vitórias de mandantes nos dados reais, então o percentual é dado por $\frac{AV_M \times 100}{TV_M}$.

Considerando os acertos dentro das tentativas, observa-se equilíbrio entre os estimadores para os percentuais de vitórias dos mandantes. Entretanto, o estimador $\hat{\theta}$ tende a prever uma quantidade um pouco inferior de vitórias de mandantes que os outros estimadores, ou seja, com um volume menor de estimativas de vitorias de mandante o estimador $\hat{\theta}$ já alcança uma quantidade de acertos bastante significativa. Trata-se portanto de um estimador mais eficiente para prever as vitórias de mandantes. Para os empates, o estimador $\hat{\theta}$ é obviamente mais eficiente, afinal os estimadores $\hat{\theta}_I$, $\hat{\theta}_{II}$ consideram muito poucos empates e o estimador $\hat{\theta}_{III}$ considera um volume excessivo de empates. Para as vitórias de visitantes, o estimador $\hat{\theta}_I$ parece mais efetivo, entretanto é possível observar se comparando $\hat{\theta}_I$ com $\hat{\theta}$, ocorre um ganho inferior à 10% nas vitórias de visitantes para $\hat{\theta}_I$, mas uma perda de quase 30% nas previsões de empates.

Avaliando os acertos gerais, para o estimador θ, fica clara uma troca entre efetividade para prever vitórias de visitantes, em relação à $\hat{\theta}_I$ (que é o melhor para este fim), para obter um ganho significativo na previsão dos empates, mas preservando a qualidade na previsão geral dos resultados. Já o estimador $\hat{\theta}_{III}$ parece muito bom para acertos em empates, entretanto isto se deve ao grande volume de empates previstos. Por outro lado, este estimador se mostra bastante ineficaz para prever vitórias, sejam elas de mandantes ou visitantes.

Dentre as 11 rodadas simuladas, destacam-se as rodadas 29, 32 e 34. O acerto do resultado usando o estimador $\hat{\theta}$ foi de 70% nas rodadas 29 e 34 e foi de 80% na rodada 32. Cabe mencionar que, dentre estas três rodadas, aproximadamente 50% dos erros ocorreram em situações cuja percentagem obtida através das simulações para o resultado estimado pelo modelo e para o resultado que realmente ocorreu eram ambas superiores à 30%, indicando assim que os resultados eram aproximadamente equiprováveis. Verifica-se tal cenário em grande parte dos erros observados, comprovando assim a natureza de difícil previsibilidade do esporte, quando o objetivo é prever pura e simplesmente o resultado de uma partida isolada. Um exemplo pode ser observado na Tabela 3 com os resultados para a trigésima segunda rodada.

As partidas Grêmio contra Bahia e Vitória contra Corinthians, na Tabela 3, ilustram bem o efeito do estimador $\hat{\theta}$, dado que o volume de empates (em verde) dentre as simulações foi superior à 107% do volume de empates ocorridos nos dados reais até a trigésima primeira rodada, o estimador considerou o resultado de empate e não a vitória do mandante como os outros estimadores sugeririam. Neste caso foram verificados dois acertos na estimativa proposta por $\hat{\theta}$, ou seja, o resultado real ocorrido nas referidas partidas foi um empate.

Tabela 3: Análise de resultados para a trigésima segunda rodada.

Partida				Probabilidades			$\hat{\theta}$ (estimativa)		
mandante	placar		visitante	V_M	E	V_V	V_M	E	V_V
Atlético-MG	5 x 0		Náutico	0,591	0,233	0,186	1	0	0
Atlético-PR	1 x 0		Internacional	0,466	0,278	0,256	1	0	0
Criciúma	1 x 1		Ponte Preta	0,304	0,295	0,401	0	0	1
Flamengo	1 x 0		Fluminense	0,456	0,257	0,287	1	0	0
Goiás	1 x 0		Botafogo	0,490	0,258	0,252	1	0	0
Grêmio	0 x 0		Bahia	0,409	0,297	0,293	0	1	0
Santos	0 x 1		Cruzeiro	0,378	0,276	0,345	1	0	0
São Paulo	2 x 1		Portuguesa	0,480	0,262	0,258	1	0	0
Vasco	2 x 1		Coritiba	0,419	0,286	0,295	1	0	0
Vitória	1 x 1		Corinthians	0,515	0,326	0,158	0	1	0

em que V_M, E, V_V representam vitória do mandante, empate e vitória do visitante respectivamente.

3.2 Premier League 2013/14

A simulação de resultados para a Premier League inglesa tem duas motivações claras. A primeira é verificar se o modelo de simulação construído se adequa bem somente ao Campeonato Brasileiro, ou se pode ser utilizado em outros certames. A segunda motivação é voltar a discussão do valor $cota = 1{,}07\varphi$, com φ sendo o percentual de empates ocorridos nos dados reais anteriores à rodada que está sendo simulada. Inicialmente as análises utilizaram a mesma cota do Campeonato Brasileiro e a mesma forneceu resultados promissores. Posteriormente, novas cotas foram testadas e a cota original se manteve efetiva frente as demais cotas avaliadas, confirmando a escolha anteriormente proposta.

Conforme já mencionado, as simulações foram realizadas apenas para partidas já realizadas. Foram simulados os resultados de jogos entre as rodadas 11 e 22. Para cada partida foram realizadas novamente 20.000 simulações. A Tabela 4 compara o volume de vitórias dos mandantes, empates e vitórias dos visitantes em dados simulados com os valores reais. Novamente trata-se de uma análise para a verificação da adequabilidade do modelo proposto com a realidade do campeonato, neste caso a Premier League. Pode-se observar que nem todos os estimadores apresentam resultados simulados que se aproximam dos valores reais. Novamente o estimador $\hat{\theta}_{III}$ não privilegiou o resultado mais esperado, a vitória do mandante.

Tabela 4: Comparação entre a ocorrências de resultados reais e simulados (volume total).

	Vitórias de Mandantes	Empates	Vitórias de Visitantes
Ocorrências Reais	53	28	39
Ocorrências Simuladas ($\hat{\theta}_I$)	81	11	28
Ocorrências Simuladas ($\hat{\theta}_{II}$)	85	0	35
Ocorrências Simuladas ($\hat{\theta}_{III}$)	38	72	10
Ocorrências Simuladas ($\hat{\theta}$)	43	42	35

A avaliação quanto à distribuição de resultados em vitórias de mandantes, visitantes e empates verificados anteriormente no Campeonato Brasileiro são confirmados para a Premier League. Os estimadores $\hat{\theta}_I$, $\hat{\theta}_{II}$ e $\hat{\theta}_{III}$ apresentaram 44%, 45% e 42% de resultados simulados iguais aos resultados reais, respectivamente. O estimador $\hat{\theta}$ apresentou 51% dos resultados simulados sendo condizentes com o resultado real da partida. A Tabela 5 apresenta para cada estimador o percentual de acerto dentro das tentativas de estimação e as percentagens de acerto em relação ao resultado real como na análise do Campeonato Brasileiro.

Considerando os acertos dentro das tentativas observa-se superioridade dos estimadores $\hat{\theta}_{III}$ e $\hat{\theta}$ em relação aos demais. Para os empates, o estimador $\hat{\theta}$ é obviamente mais eficiente, afinal os estimadores $\hat{\theta}_I$, $\hat{\theta}_{II}$ consideram muito poucos empates e o estimador $\hat{\theta}_{III}$ considera um volume

Tabela 5: Análise do número de resultados simulados corretamente.

Acerto dentro das tentativas			
Estimador	Vitórias de Mandantes	Empates	Vitórias de Visitantes
$\hat{\theta}_I$	46,9%	18,2%	46,4%
$\hat{\theta}_{II}$	45,9%	0,0%	42,9%
$\hat{\theta}_{III}$	65,8%	29,2%	40,0%
$\hat{\theta}$	58,1%	38,1%	42,9%
Acerto geral			
Estimador	Vitórias de Mandantes	Empates	Vitórias de Visitantes
$\hat{\theta}_I$	71,7%	7,1%	33,3%
$\hat{\theta}_{II}$	73,6%	0,0%	38,5%
$\hat{\theta}_{III}$	47,2%	75,0%	10,3%
$\hat{\theta}$	47,2%	57,1%	38,5%

excessivo de empates. Para as vitórias de visitantes os quatro estimadores apresentam resultados semelhantes e em todos os casos com um percentual bom. O estimador $\hat{\theta}_I$ novamente parece mais efetivo, entretanto é possível observar se comparando $\hat{\theta}_I$ com $\hat{\theta}$, ocorre um ganho inferior à 5% nas vitórias de visitantes a favor de $\hat{\theta}_I$, mas uma perda de quase 20% nas previsões de empates e de 10% nas vitórias de mandantes.

Avaliando os acertos gerais, para o estimador θ novamente é possível observar uma troca entre efetividade para prever vitórias de visitantes em relação à $\hat{\theta}_I$ e $\hat{\theta}_{II}$, para obter um ganho significativo na previsão dos empates, mas preservando a qualidade na previsão geral dos resultados. Já o estimador $\hat{\theta}_{III}$ parece muito bom para acertos em empates (devido ao grande volume de empates previstos) e ainda adequado para vitórias de mandantes, entretanto se mostra bastante ineficaz para prever vitórias de visitantes. O ganho em acertos de vitórias de visitantes dos estimadores $\hat{\theta}_I$ e $\hat{\theta}_{II}$ em relação a $\hat{\theta}$ é de aproximadamente 25%, porém a perda nos acertos de empates é da ordem de 50%.

Dentre as 12 rodadas simuladas, destacam-se as rodadas 16 e 19. O acerto do resultado usando o estimador $\hat{\theta}$ foi de 80% na rodada 19 e foi de 90% na rodada 16. Vale ressaltar que, dentre estas duas rodadas, dos três erros cometidos, dois eram casos em que a percentagem obtida através das simulações para o resultado estimado pelo modelo e para o resultado que realmente ocorreu eram ambas superiores à 30%, indicando assim que os resultados eram aproximadamente equiprováveis. Verifica-se novamente tal cenário em grande parte dos erros observados. Um exemplo pode ser observado na Tabela 6 com os resultados para a décima sexta rodada.

Tabela 6: Análise de resultados para a décima sexta rodada.

Partida				Probabilidades			$\hat{\theta}$ (estimativa)		
mandante	placar		visitante	V_M	E	V_V	V_M	E	V_V
Aston Villa	0 x 3		Manchester United	0,362	0,251	0,388	0	0	1
Cardiff	1 x 0		West Bromwich	0,342	0,267	0,391	0	0	1
Chelsea	2 x 1		Crystal Palace	0,634	0,216	0,150	1	0	0
Everton	4 x 1		Fulham	0,659	0,186	0,155	1	0	0
Hull City	0 x 0		Stoke City	0,540	0,236	0,224	0	1	0
Manchester City	6 x 3		Arsenal	0,689	0,159	0,152	1	0	0
Newcastle	1 x 1		Southampton	0,455	0,288	0,257	0	1	0
Norwich City	1 x 1		Swansea City	0,398	0,298	0,304	0	1	0
Tottenham	0 x 5		Liverpool	0,362	0,243	0,395	0	0	1
West Ham	0 x 0		Sunderland	0,545	0,294	0,161	0	1	0

em que V_M, E, V_V representam vitória do mandante, empate e vitória do visitante respectivamente.

Novamente na Tabela 6 o efeito do estimador estimador $\hat{\theta}$ transformando alguns casos (em verde) em que $\max(V_M, E, V_V) = V_M$ em empates levou a proposta de estimativa para quatro acertos de resultados que não seriam obtidos nos estimadores anteriores.

4 Conclusões

Não existe grande novidade ao mencionar o alto grau de complexidade associado ao problema de previsão de resultados esportivos, em particular para as partidas de futebol. Considerando a simplicidade do modelo proposto, que leva em consideração apenas a pontuação básica das partidas já realizadas, a proposta aqui discutida apresenta um conjunto de resultados bastante promissores.

A distribuição Poisson truncada à direita se mostrou eficiente para mensurar o efeito probabilístico da variável aleatória número de gols de uma equipe em uma dada partida de futebol. O estimador proposto para a taxa λ associada à distribuição Poisson é inovador em sua técnica de reduzir a superestimação para a probabilidade associada ao valor 0.

O modelo apresentou resultados condizentes com os volumes de vitórias de mandantes, visitantes e empates não somente no Campeonato Brasileiro da série A, mas também na Premier League inglesa. A qualidade deste resultados é obviamente fruto da proposição da cota para determinação de empates imposta ao estimador de resultados de partidas. Esta cota associada ao volume de empates já ocorridos no campeonato se adaptou bem aos dois campeonatos avaliados e espera-se que o comportamento se repita na avaliação de outros certames. Em muitas situações ilustradas na seção 3 é possível verificar o efeito da cota corrigindo resultados que um estimador mais ingênuo não seria capaz de fazê-lo.

Mediante aos estudos apresentados, a possibilidade de previsão de resultados de campeonatos completos de futebol deixa de permear apenas o campo das idéias, tornando-se extremamente plausível e bem fundamentada através do estimador proposto. Consolidado o estimador para previsão de partidas, a previsão de torneios completos se torna, a menos de possíveis ajustes e adequações, uma questão apenas computacional. Uma outra possibilidade a ser avaliada é a previsão dos placares de cada partida e não somente prevendo vencedores e empates. Obviamente trata-se de um problema mais sofisticado, devido à magnitude de seu espaço de possíveis resultados, cujo tamanho supera na maior parte das vezes o tamanho do espaço avaliado quando considerando apenas vitórias e empates, espaço este composto por apenas três possíveis resultados.

A continuidade deste estudo prevê o desenvolvimento das técnicas necessárias à previsão de resultados para os campeonatos completos, aplicação em torneios já encerrados para verificação de sua capacidade preditiva, utilização em torneios ainda em andamento e ainda a possível verificação dos placares das partidas por meio de simulação, e a verficação da qualidade dos placares previstos.

Referências

[1] BRILLINGER, D. R. An analysis of Chinese Super League partial results. **Science in China Series A - Mathematics**. v. 52, p. 1139-1156, 2009.

[2] FARIAS, F. F. **Análise e Previsão de Resultados de Partidas de Futebol**. 2008. 78 p. Dissertação (Mestrado em Estatística), Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2008.

[3] KARLIS, D.; NTZOUFRAS, I. Analysis of sports data by using bivariate Poisson models. **Journal of the Royal Statistical Society. Series D (The Statistician)**. v. 52, p. 381-393, 2003.

[4] KARLIS, D.; NTZOUFRAS, I. On model Soccer Data. **Student**. v. 3 (4), p. 229-244, 2000.

[5] KNORR-HELD, L. Dynamic Rating of Sports Teams. **Journal of the Royal Statistical Society. Series D (The Statistician)**. v.49 (2), p. 261-276, 2000.

[6] FANG, L.; ZHENG, Z. Predicting Soccer League Games using Multinomial Logistic Models. **Relatório Técnico**, 2008.

[7] RUE, H.; SALVESEN, Ø. Prediction and Retrospective Analysis of Soccer Matches in a League. **Journal of the Royal Statistical Society. Series D (The Statistician)**, v. 49 (3), p. 399-418, 2000.

[8] SOUZA JR., O. G.; GAMERMAN, D.. Previsão de Partidas de Futebol usando Modelos Dinâmicos. In: XXXVI SBPO - Sociedade Brasileiro da Pesquisa Operacional. **Anais do XXXVI SBPO**, 2004.

[9] SUPERESPORTES. Disponível em: `http://www.superesportes.com.br/`, acesso em: 20 de Janeiro de 2014.

[10] Tabela do Brasileirão. Disponível em: `http://www.tabeladobrasileirao.net/`, acesso em: 20 de Janeiro de 2014.