UFSM

CIÊNCIA e NATURA

ISSN 2179-460X
ACESSO ABERTO

Ci. e Nat., Santa Maria, v. 45, e3, 2023 • https://doi.org/10.5902/2179460X68508
Submissão: 16/11/2021 • Aprovação: 22/09/2022 • Publicação: 31/01/2023

Estatística

Proposta de uma métrica para o efeito da casa baseada em pontos ganhos

A propose of a metric to the home advantage based on awarded points

Giovani Festa Paludo[I], **Nikolas Neves de Figueiredo**[I], **Eric Batista Ferreira**[I]

[I] Universidade Federal de Alfenas, Alfenas, MG, Brasil

ABSTRACT

Em competições esportivas de pontos corridos, o total de pontos ganhos em casa sobre o total de pontos consiste em uma métrica para medir a vantagem da casa que é fácil de utilizar, de interpretar e que é muito conhecida. Porém, quando um time faz poucos pontos essa métrica fica inflacionada e oscila expressivamente, entre outras fragilidades. Assim, nosso objetivo foi construir uma nova métrica para a vantagem da casa mantendo as características consideradas como positivas dessa métrica e ao mesmo tempo superando as fragilidades listadas. Com isso, foram propostos uma métrica e um teste para serem utilizados em campeonatos esportivos. Como aplicação, foi utilizado o Brasileirão Série-A (2003-2020). Como principal resultado, a nova métrica não foi negativamente correlacionada com os pontos conquistados. O teste sobre a métrica permitiu verificar quando há informação suficiente para afirmar que existe vantagem da casa, tirando o efeito da oscilação quando o time ganha poucos pontos. Com base nisso, das 370 participações de 43 times em 18 edições, uma apresentou efeito da casa negativo, 259 tiveram efeito da casa positivo e 110 não apresentaram efeito da casa. Ainda, a métrica permitiu visualizar o comportamento individual de cada time ao longo das edições da competição.

Keywords: Efeito mandante; Estatísticas do esporte; Futebol; Campeonato Brasileiro; Liga

RESUMO

In double round-robin sports tournaments, the total of points awarded divided by the total of the points earned at home consists in metric to access the home advantage that is easy to use, to interpret and is widely known. However, when a team award a small number of points this metric is underestimated, vary expressively, among others fragilities. Thus, our objective was to build a new metric to measure home advantage keeping the positive aspects of the previous one and improving all the listed fragilities. It was proposed the new metric and a test to be used in sportive championships. As the study application, it was used data from Brazilian Championship Series-A (2003-2020). As our main result, the new metric

Artigo publicado por Ciência e Natura sob uma licença CC BY-NC-SA 4.0.

was not negatively correlated to the awarded points. The test of the metric let to verify on which occasion there is enough information to affirm that exists home advantage, avoiding the effects of a small number of awarded points. Based on this, of the 370 participations of 43 teams in 18 editions, one participation presented home disadvantage, 259 presented home advantage and 110 did not present home effect. Besides, the metric permitted to visualize the individual behavior of each team throughout the league.

Palavras-chave: Home win; Home ground effect; Home cocking; Soccer; Sports statistics

1 INTRODUÇÃO

Quando o time do coração vai jogar em casa, o torcedor já cria expectativas de que o time tem maiores chances de ganhar. E o torcedor não está errado, pois é bem conhecido no mundo dos esportes, o fato de que a probabilidade de um time obter uma vitória em uma partida é maior quando o time joga em casa (POLLARD; POLLARD, 2005). A esse fenômeno, de melhores resultados jogando em casa, é comumente atribuído o termo "vantagem da casa" (HA). Sendo que HA pode ser definida como "o resultado consistente no qual o time da casa em competições esportivas vence mais que 50% dos jogos em um calendário de jogos balanceado entre casa e fora de casa" (COURNEYA; CARRON, 1992, p. 13). Ainda, a HA pode ser encontrada em diferentes esportes (POLLARD; POLLARD, 2005; DAWSON *et al*., 2020), é persistente ao longo do tempo (POLLARD; POLLARD, 2005; JACKLIN, 2005) e existem diferentes métricas para obtê-la (POLLARD *et al*., 2008; GOUMAS, 2017; MAREK; VÁVRA, 2020).

Em competições esportivas no formato de pontos corridos, medir a vantagem da casa por pontos é bastante útil, pois conquistar pontos é o principal objetivo de cada time nesse tipo de competição. Por isso, entre as métricas utilizadas para obtenção de HA, as baseadas em pontos são frequentemente utilizadas (POLLARD *et al*., 2008; LEITE, 2017; OLIVEIRA *et al*., 2020) e destaca-se a métrica obtida a partir da divisão entre o total de pontos conquistados em casa e o total de pontos conquistados, que neste estudo será chamada de métrica *h* e que já foi utilizada em vários estudos com ou sem alguma correção (FAJARDO *et al*., 2019; TILP; THALLER, 2020; POLLARD *et al*., 2008) (sendo que

h assume valores entre 0 e 100%, em que *h* = 50% é sem vantagem da casa, *h* = 100% é interpretado como o máximo de HA e *h* = 0% seria o máximo de desvantagem da casa). A métrica *h* é baseada em pontos, é fácil de ser calculada e interpretada, porém, é possível apontar algumas fragilidades.

A primeira fragilidade para se obter adequadamente um valor da função *h* para um time acontece quando o time conquista poucos ou muitos pontos do total de disputados. Para um exemplo, considere uma competição com as seguintes características: formato de pontos corridos, onde todos os times se enfrentam entre si duas vezes, uma vez na casa do primeiro e outra na casa do segundo; com 20 times; com 38 rodadas e; onde o vencedor de uma partida ganha 3 pontos e no caso de empate fica 1 ponto para cada time. Então, dentro dessa liga, considere o time A e o time B (Tabela 1) que são exemplos hipotéticos gerados para evidenciar as fragilidades. Como fez poucos pontos, o time A pode ter valores de *h* de 0% a 100%, mas o time B que fez muitos pontos só pode ter a métrica *h* de 46% a 54%, assim a métrica *h* só permite valores de 100% quando o time pontua de 1 a até a metade do número total de partidas. Com base nisso, é possível observar que o valor de *h* depende dos pontos conquistados e provavelmente este é o motivo pelo qual pesquisas utilizaram metodologias para corrigir o valor de *h* em relação aos pontos conquistados (frequentemente chamada de correção pela habilidade) (CLARKE; NORMAN, 1995; POLLARD *et al*., 2008; GOUMAS, 2017; VAN-OURS, 2019). Porém, no próximo parágrafo será pontuado porque uma correção baseada apenas em habilidade pode desconsiderar outra informação que é pertinente.

Tabela 1- Exemplo com 6 times hipotéticos (A, B, C, D, E e F) e dados hipotéticos ao final de uma edição do campeonato para mostrar as fragilidades da métrica *h*. Sendo que "Vit." significa vitória, "Emp." significa empate e "Der." significa derrota. Já "Números totais" se refere ao total de jogos em casa e fora de casa

	Números em casa			**Números totais**				
Time	**Vit.**	**Emp.**	**Der.**	**Vit.**	**Emp.**	**Der.**	**Pontos**	***h***
A	-	-	-	3	0	35	9	De 0% a 100%
B	-	-	-	35	0	3	105	De 46% a 54%
C	-	-	-	0	33	5	33	De 42% a 58%
D	-	-	-	11	0	27	33	De 0% a 100%
E	2	0	17	3	0	35	9	67%
F	1	0	18	3	0	35	9	33%

Fonte: Autores (2022)

A segunda fragilidade a ser destacada é o fato de que o número de pontos conquistados é importante, mas o número de vitórias e o número de empates é mais importante que o número de pontos em si. Observe os times C e D (Tabela 1). Ambos têm a mesma pontuação, só que D ganhou os pontos em vitórias e C ganhou os pontos em empates. O time C, matematicamente falando, poderá ter o valor de *h* entre 42% e 58%, enquanto que matematicamente falando, o time D poderá ter *h* de 0% a até 100%. Esse exemplo mostra que para realizar a correção no valor da métrica *h* deve-se considerar os pontos obtidos por empate e por vitória. Sendo que, embora hajam correções para a métrica *h*, os autores do presente estudo desconhecem algum estudo que tenha utilizado uma correção que considera o número de empates ou de vitórias.

A terceira fragilidade está relacionada a situações em que um time ganha poucos pontos. Seguindo o mesmo formato de liga acima descrita e considerando como exemplo os times E e F (Tabela 1). Ambos tiveram apenas 3 vitórias em um campeonato, porém o time E obteve uma vitória a mais em casa que o time F. Com isso, a *h* para o time E seria de 67% e para o time F seria 33%. A diferença de uma vitória de um total de 3 vitórias é suficiente para afirmar que o E tenha 67% e o time F 33%? Certamente não é suficiente. Ainda, fica evidenciado a oscilação que um número pequeno de vitórias provoca no valor da métrica. E por último, o time E e F podem ter tido essa diferença de 1 vitória por um efeito aleatório ou não relacionado, pois uma partida de diferença pode não ser suficiente para determinar que a *h* do time E foi 67% e que a *h* do F foi 33%.

Associado à essa segunda fragilidade, podemos observar que os estudos nem sempre trazem informações de vantagem da casa para todos os times participantes. Eventualmente, é possível observar uma exclusão de certos times que participaram poucas vezes, por conta de uma limitação da metodologia empregada. Por exemplo, Goumas (2017) analisou uma competição no formato copa e mostrou os valores apenas para os times que tiveram mais de 50 partidas. Pollard *et al*. (2008) calcularam valores de vantagem da casa apenas para times que participaram no mínimo em 3 edições do campeonato em formato de liga. Acredita-se no presente estudo que se um time participou uma única vez, isso é suficiente para que ele tenha alguma informação sobre o efeito da casa. E esta é a questão central desta pesquisa, estabelecer uma métrica por pontos e sem as fragilidades acima listadas. Uma métrica sem essas limitações contribuirá para investigações do que está associado à vantagem da casa, sendo que permitirá que futuros estudos utilizem um menor tamanho amostral e, ainda, a métrica permitirá a obtenção de dados longitudinais, o que contribuirá para o melhor entendimento do comportamento da vantagem de se jogar em casa nos times ao longo do tempo.

Considerando competições de esportes coletivos nas quais são jogadas o mesmo número de partidas em casa e fora de casa, o objetivo do presente estudo foi de desenvolver uma métrica para medir o efeito da casa que mantém as características positivas da métrica *h*, tais como: ser baseada em pontos conquistados, ser de fácil obtenção e interpretação. E também superar as três fragilidades e a observação apontadas acima, isto é: (i) espera-se que a nova métrica não é inflacionada quando o time obtém poucos pontos. Se essa afirmação é verdadeira, será confirmado que: a vantagem da casa obtida pela métrica *h* deverá ser negativamente correlacionada com os pontos conquistados, enquanto que a vantagem da casa obtida pela nova métrica não será negativamente correlacionada com os pontos conquistados. (ii) que avalie quando há evidências suficientes para concluir que exista efeito da casa e não seja apenas um efeito devido ao acaso. Ou seja, para superar a segunda e a terceira fragilidade também foi objetivo apresentar um teste para avaliar quando se tem evidências suficientes para concluir ou não que é um efeito da casa, e não apenas um efeito devido a questões ao acaso como pouca informação. Sendo que ainda o teste deve considerar a configuração de vitórias e empates que aconteceu ao invés de considerar apenas os pontos.

2 MATERIAL E MÉTODOS

2.1 Descrição dos dados

Como exemplo de aplicação foi utilizado o Campeonato Brasileiro de Futebol Série A, utilizando-se dados de todas as edições da história dos pontos corridos, ou seja, de 2003 a 2020. A partir de 2003, todos os times passaram a se enfrentar duas vezes e o campeão passou a ser o time que acumulou mais pontos somando-se todos os enfrentamentos. Sendo que em 2003 e 2004, o campeonato tinha 24 times participantes, 46 rodadas e 552 partidas. Já em 2005, foram 22 times, 42 rodadas e 462 partidas. No de 2006 até os dias atuais o campeonato tem 20 times participantes, 380 partidas em 38 rodadas. Os

dados utilizados no presente estudo foram obtidos no www.soccerway.com, website que também foi utilizado nos estudos de Pollard *et al*. (2008); Silva *et al*. (2018). As variáveis coletadas foram: nome do time da casa, nome do time visitante, gols do time da casa e gols do time visitante.

Ainda, há algumas considerações sobre dados não considerados no presente estudo: (i) como aqui foi utilizado a pontuação conquistada por partida, então todas as punições que implicaram na subtração ou adição de pontos na tabela de pontuação do campeonato não foram consideradas no presente estudo. Tais subtrações ou adições aconteceram em 2003, 2004, 2005, 2010 e 2013 (2003: -4 Ponte Preta, -8 Paysandu, +3 São Caetano, +3 Ponte Preta, +2 Corinthians, +2 Fluminense, +3 Juventude e +2 Internacional; 2004: -24 São Caetano; 2005: -1 Brasiliense e +2 Vasco; 2010: -3 Grê. Barueri; 2013: -4 Portuguesa e -4 Flamengo); (ii) há 23 partidas ocorridas nos anos de 2016, 2018 e 2019 nos quais houve venda do mando de campo referencial por motivos distintos. Com isso, a partida ocorreu em outro local, sendo que essas partidas não foram removidas do banco de dados; (iii) não foram consideradas as partidas anuladas no ano de 2005, sendo que foram consideradas apenas as novas partidas e; (iv) a pontuação da partida da Chapecoense na última rodada da edição de 2016 foi considerada de maneira igual à considerada pela Confederação, que foi de derrota para ambos os times.

Cabe ainda ressaltar que, por exemplo, em uma partida entre Grêmio *versus* Vasco. Entendeu-se o primeiro time (Grêmio) como mandante e o segundo time como visitante (Vasco), não importando o estádio em que o jogo aconteceu. Assim, todas as vezes que os termos “partida em casa” e “partida fora de casa” foram utilizados, eles se referiram, respectivamente à “partida como mandante” e “partida como visitante”.

2.2 Métrica Proposta e Análise de Dados

Para medir o quanto o desempenho de um time é melhor em casa do que como visitante, foi necessário estabelecer algumas definições. A primeira delas foi

a diferença de pontos absoluta (*da*) que ficou definida como a diferença entre os pontos conquistados em casa (P_c) e os pontos conquistados fora de casa (P_f) no final do campeonato, isto é:

$$da = P_c - P_f .$$

Em seguida, estabeleceu-se a diferença de pontos relativa (*d*), que ficou definida como a diferença de pontos absoluta multiplicada por um fator que faça com que o maior valor possível de diferença, seja 100%, isto é,

$$d = \frac{P_c - P_f}{m \times v} \times 100, \tag{1}$$

em que *v* são os pontos por vitória; *m* é o número de partidas ou rodadas que um time joga em casa em uma edição do campeonato.

Ainda, o efeito da casa médio amostral de um time ($\bar{D}$) obtido pela métrica *d*, será definido como a média dos valores de *d*, isto é:

$$\bar{D} = \frac{\sum_{i=1}^{n} d_i}{n},$$

em que *i* é cada participação do time em uma edição do campeonato e *n* é o número total de vezes que um time participou. Sendo que $\bar{H}$ e $\overline{DA}$ são médias obtidas da mesma forma que $\bar{D}$, isto é, $\bar{DA} = \sum_{i=1}^{n} da_i/n$ e $\bar{H} = \sum_{i=1}^{n} h_i/n$. Ressalta-se que tanto o $\bar{H}$, $\overline{DA}$ e $\bar{D}$ são médias amostrais obtidas de um número finito de participações de um espaço amostral com infinitas participações possíveis. Considerando a métrica *d*, porém, pode-se obter uma média amostral $\bar{D}$, dado a impossibilidade de obtenção da média populacional δ, e ainda pode-se obter o intervalo de confiança para a média populacional (*IC*[δ; 95%]). O mesmo acontece para a métrica *da*, sendo que a média populacional ξ é desconhecida, sendo que o que pode ser obtido é uma média amostral $\overline{DA}$, e um intervalo de confiança para a média populacional *(IC[ξ; 95%*]). Para a obtenção

das estimativas intervalares para a média populacional foi utilizado a estatística *t* a 95% de probabilidade do intervalo conter o verdadeiro parâmetro populacional.

Uma maneira bastante comum para se medir o efeito positivo da casa é a métrica *h* sem correção, que diz respeito à porcentagem dos pontos conquistados em casa em relação ao total de pontos conquistados. Podendo ser escrita na forma de

$$h = \frac{P_c}{P_c + P_f} \times 100, \tag{2}$$

em que: P_c é a variável pontos conquistados em casa; P_f é a variável pontos conquistados como visitante e; o valor de h = 100% indica o máximo de vantagem da casa, h = 50% indica nenhuma vantagem da casa e h = 0% indicaria o máximo de desvantagem da casa.

Para permitir a comparação da estatística criada no presente estudo com a estatística *h* (2) já desenvolvida, foi utilizada a vantagem da casa centrada em 0 (h_0), que ficou definida como:

$$h_0 = 2h - 100,$$

em que h_0 varia de −100% quando a desvantagem da casa é total, 0 quando não há vantagem e nem desvantagem e h = 100% quando a vantagem da casa é total. Cabe ressaltar que o h e h_0 do presente artigo são as mesmas métricas apresentadas em estudo de Matos *et al*. (2020), onde foram denominados de tradicional e reescalado, respectivamente.

Já a média global de todos os valores *d* foi obtida pela média aritmética dos 370 valores de *d*, sendo que cada valor de *d* correspondeu a uma participação de um time em uma edição da competição. É possível obter um *d* exato para toda a competição, porém não foi utilizado dentro do presente artigo, mas foi apresentado separadamente na Apêndice A. Ressalta-se aqui, que d, h, h_0 ficaram estabelecidos como as métricas acima definidas. Por outro lado, HA não é uma métrica nem uma função, sendo que HA foi utilizado como uma abreviatura de "vantagem da casa" ou de "efeito positivo da casa" ou "efeito mandante positivo".

2.3 Inferência para uma participação

Como exposto na terceira fragilidade, há diferentes valores possíveis de vantagem da casa dependendo se o ponto veio de um empate ou vitória. Então, inicialmente foi escrito o vetor denominado *P* que foi composto por todos os valores de pontos obtidos por partida de um certo time em uma edição do campeonato, indo de 1 a 38 no Campeonato Brasileiro.

$$P = \{p_1, p_2, p_3, \ldots, p_{38}\},$$

em que p_1 é a pontuação obtida na partida da rodada 1, p_2 na rodada 2, e assim por diante.

Com isso foi definido o vetor *R*, um novo vetor constituído com os mesmos valores do vetor *P*, porém com os valores em uma ordem diferente da original. Sendo que a ordem foi obtida com um processo de aleatorização, e assim foi escrito um novo vetor:

$$R = \{r_1, r_2, r_3, \ldots, r_{38}\}.$$

Para gerar uma distribuição com os possíveis valores da métrica *d*, foi então definido que os pontos em casa seriam a primeira metade dos elementos e os pontos conquistados fora de casa a segunda metade do conjunto.

$$P_c = \sum_{i=1}^{19} r_i \text{ e } P_f = \sum_{i=20}^{38} r_i$$

O próximo passo constituiu-se na obtenção do conjunto com os valores de *d*. Ou seja, aplicou-se a expressão (1) para cada sorteio e obteve-se um valor de *d* que foi chamado de ds_i, onde *i* é o número da simulação, e *ds* é a diferença de pontos relativa de cada simulação. Ao final de 1000000 de simulações, obteve-se o conjunto:

$$DS = \{ds_1, ds_2, \ldots, ds_{1.000.000}\}$$

Em seguida ordenou-se esse conjunto e foram obtidos os percentis de interesse,

que no caso foi deixando 20% dos valores nas caudas, isto é, os percentis de 10% e 90%. Sendo que este valor de 20% nas caudas foi escolhido pois não é rara a situação que um time termina o campeonato com poucos pontos, e considerou-se como melhor alternativa detectar o efeito da casa mesmo que haja pouca informação. Assim definimos o intervalo de predição (*IP*) composto pelo limite superior (IP_{LS}) e pelo limite inferior (IP_{LI}) que podem ser escritos da forma:

$$IP_{LS} = Sort(DS)_{0.90*k} \text{ e } IP_{LI} = Sort(DS)_{0.10*k},$$

em que *Sort* é uma função do Software R que ordena os elementos do conjunto *DS* em ordem crescente e *k* é o número de reamostras.

A verificação da existência do efeito da casa não nulo foi realizada pela comparação entre a métrica *d* e os IP_{LS} e IP_{LI}. Admite-se neste trabalho que, se $d < IP_{LI}$, então há efeito da casa negativo que pode ser chamado de desvantagem da casa; se $IP_{LI} \leq d \leq IP_{LS}$, então não há evidencias suficientes para afirmar que há efeito da casa e; se $IP_{LS} < d$, então há efeito da casa positivo, ou seja, há vantagem da casa.

Ainda, para verificar se existia correlação entre *h* e os pontos obtidos e entre *d* e os pontos obtidos, foi utilizada uma correlação linear simples e um teste *t* para avaliar a significância do coeficiente *r* de correlação de linear simples. E nessas duas correlações foram utilizadas as edições com 20 times participantes, ou seja, todas as 15 edições que ocorreram de 2006 a 2020. Todas as análises e gráficos foram realizadas utilizando o software estatístico R (R Core Team, 2021), sendo que foi incluído um algoritmo escrito em linguagem R para a obtenção da métrica *d*, *da*, IP_{LI} e IP_{LS} no Apêndice B deste artigo.

3 RESULTADOS E DISCUSSÃO

3.1 Características e média global da métrica *d*

A Série A do Campeonato Brasileiro de Futebol de 2003 a 2020 teve 12879 pontos conquistados em casa e 7014 pontos conquistados fora de casa. Isso gerou uma diferença

absoluta global de pontos de *da* = 5865 pontos. A média global de todos os 370 valores obtidos da métrica *d* foi 26,80% e quando a métrica *h* foi utilizada foi obtida uma média global de 64,74%, que resultou em um h_0 global de 29,48%. Três aspectos podem ser observados em relação à métrica *d* nesses três resultados.

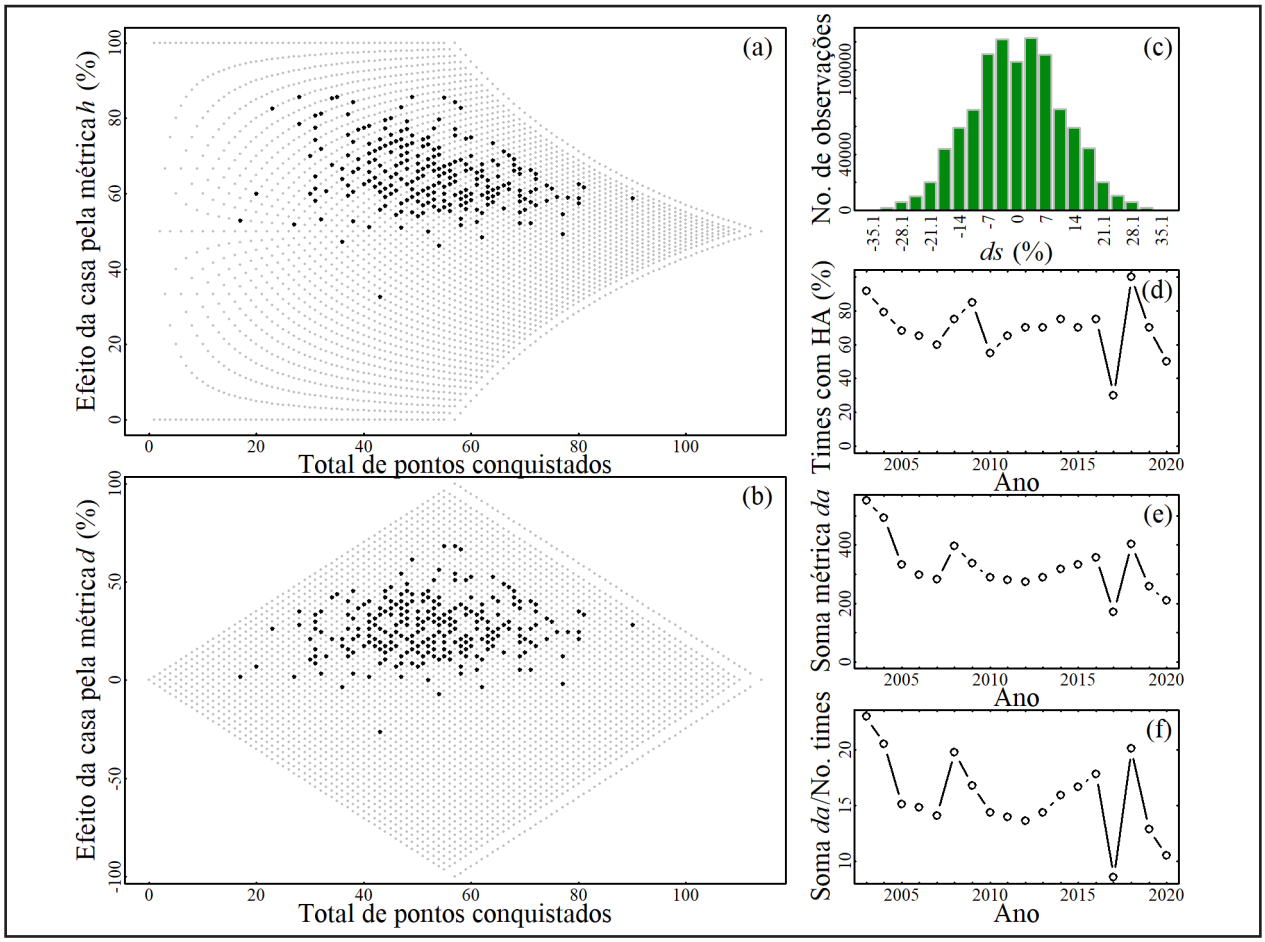
O primeiro aspecto é sobre sua interpretabilidade. Como a métrica *d* pode receber valores de, no mínimo, -100% quando o time conquistou todos os pontos possíveis fora de casa e nenhum em casa, 100% (valor máximo) quando o time conquistou todos os pontos possíveis em casa e nenhum fora de casa e, *d* é *0* quando o mesmo número de pontos foi conquistado em casa ou fora de casa. Então, a média 26,80% de todas os valores de *d* significa que, do total de pontos disputados, em média 26,80% foram conquistados a mais em casa do que como visitante. Sendo que *d* é uma métrica que reflete diretamente a porcentagem dos pontos ganhos a mais em casa do que fora de casa. Essa porcentagem pode ser utilizada para comparação de competições de pontos corridos de diferentes esportes que utilizem um sistema semelhante de pontuação.

Um segundo aspecto a ser destacado é que os valores das métricas *d* e *h* são valores próximos e podem ser comparados desde que *h* seja multiplicado por dois e subtraído em 100 unidades, isto é, que seja obtido o h_0. Esta proximidade pode ser observada em pelo menos dois gráficos (na Figura 1a comparada com a Figura 1b ou na Figura 2) que mostram tanto os valores de *d* quanto os valores de h_0 por time e por ano. Esse valor médio da métrica *d* encontrado no presente trabalho está próximo ao encontrado na literatura. Pollard *et al*. (2008) estudaram as edições de 2003 a 2007 e utilizaram a mesma métrica *h*, porém com uma correção, sendo que encontraram um *h* médio de 65% de vantagem da casa que gera um h_0 médio global de 30%. Já Fajardo *et al*. (2019) encontraram a média da métrica *h* de 65,6% de 2012 a 2016, que gera um h_0 global de 31,2%. Por outro lado, Oliveira *et al*. (2020) estudaram a edição de 2017 utilizando uma métrica diferente das utilizadas no presente artigo e observaram 54% de vantagem da casa. No presente estudo, a média de *d* para quem participou em 2017 foi de 15,0%, a menor média registrada na sequência de 2003 a 2020. Assim, um característica da métrica *d* é que a métrica corrige as fragilidades

supracitadas e ao mesmo tempo traz um valor não muito distante ao que foi obtido pela métrica *h* em estudos anteriores também sobre o Campeonato Brasileiro, mesmo que tenham ou não utilizado alguma correção (POLLARD *et al*., 2008; FAJARDO *et al*., 2019).

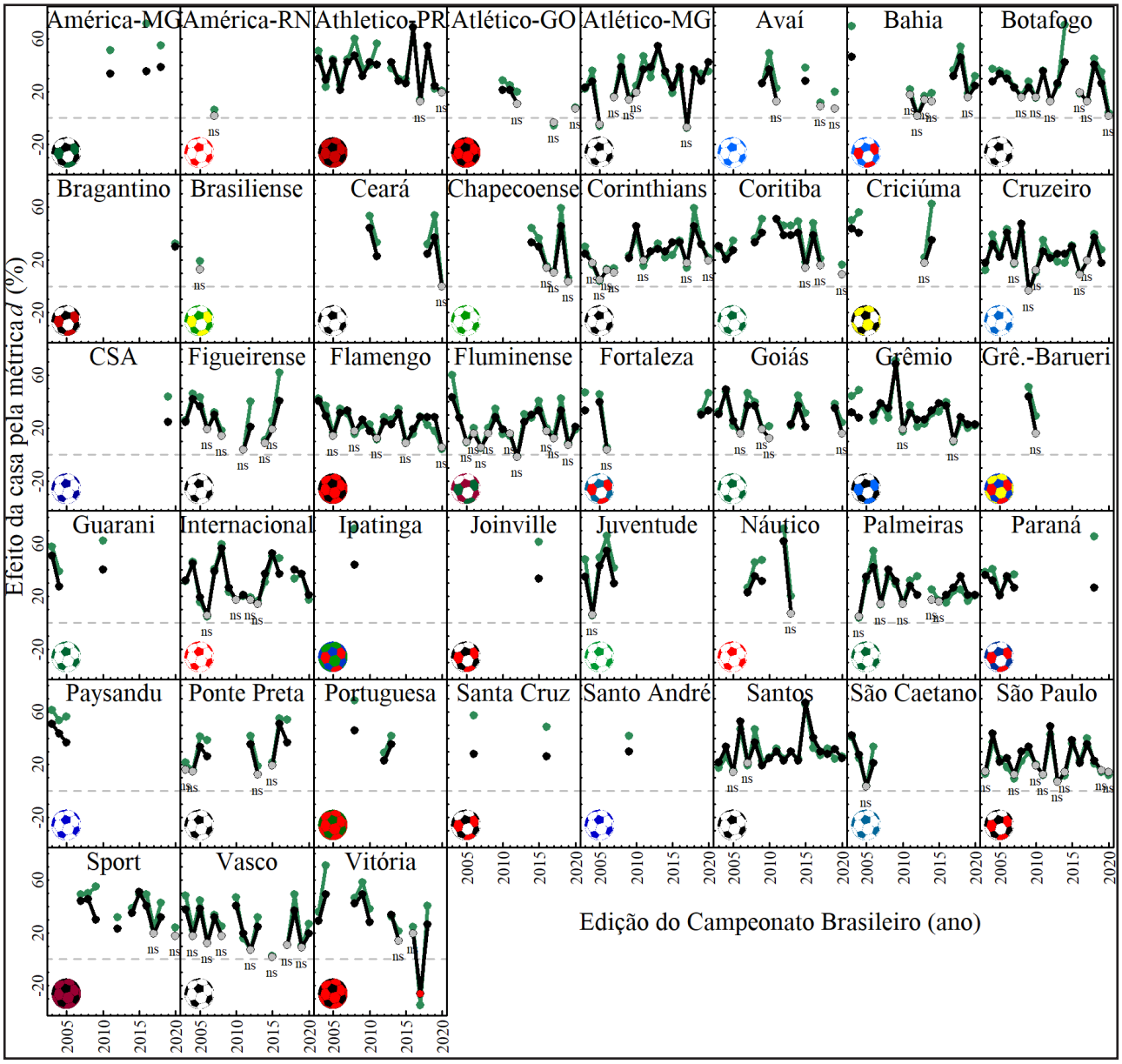
Um terceiro aspecto importante a ser destacado é que a principal diferença entre os valores de *d* e h_0 obtidos para o Campeonato Brasileiro, se deve principalmente à natureza da métrica *h*, que, ao dividir a diferença de pontos pelo total de pontos, acaba inflacionando a métrica, pois quanto menor o número de pontos, o valor da métrica tenderá a ficar maior. A métrica *d* não faz essa divisão e como é possível observar nos pontos em cinza na Figura 1a e 1b, os possíveis valores da métrica *d* são simétricos, enquanto que os possíveis valores de *h* são assimétricos quando observados ao longo dos pontos totais conquistados. Esse resultado é corroborado pela análise de correlação linear simples entre o total de pontos conquistados de cada time e o respectivo valor da métrica *h* (2) que mostrou uma correlação linear negativa ($r = -0{,}296$; $n = 300$; valor-p < 0,0001; Figura 1a). Ao contrário, a análise de correlação linear simples entre os pontos conquistados de cada time em cada ano e o valor da métrica *d* (1), não apresentou correlação significativa ($r = 0{,}101$; $n = 300$; valor-p = 0,081; Figura 1b). Assim, confirmou-se o que era esperado que a vantagem da casa obtida pela nova métrica *d* não foi negativamente correlacionada com os pontos ganhos. Com isso, o resultado suporta que a nova métrica não é inflacionada quando se tem situações de poucos pontos, isto é, a métrica não é inflacionada pela habilidade do time. Vários estudos sobre o efeito positivo da casa fazem uma correção pelos pontos conquistados e tal correção não é recente (CLARKE; NORMAN, 1995; POLLARD; GÓMEZ, 2007; POLLARD *et al*., 2008). Assim, ao contrário do que acontece com a métrica *h* sem correção, a métrica *d* não foi inflacionada quando um time fez poucos pontos, não sendo mais necessário realizar uma correção pela habilidade. Característica importante para uma métrica para acessar a vantagem da casa baseada em pontos. Ainda, a eliminação da inflação pode contribuir também com estudos que fazem previsões de resultados de jogos de futebol (ex. Ramos *et al*. (2021)).

Figura 1 – (a) O efeito da casa obtido pela métrica *h* para 300 participações em 15 edições (2006 a 2020) do Campeonato Brasileiro (pontos pretos) e todos os valores possíveis para a métrica *h* (pontos cinza); (b) O efeito da casa obtida pela nova métrica *d* para as 300 participações de 2006 a 2020 do campeonato brasileiro (pontos pretos) e todos os valores possíveis para a métrica *d* (pontos cinza). (c) Distribuição dos valores de diferença de pontos relativa simulada obtidas nas 1000000 reamostras para a configuração de vitórias, empates e derrotas do Naútico em 2013. Essa distribuição empírica indica quais são os valores mais prováveis para serem encontrados em um novo ensaio, dado o que já aconteceu. (d) Número de times que tiveram efeito da casa positivo em relação ao total de times que participou da edição do campeonato para as 18 edições de 2003 a 2020. (e) Soma da métrica *da* para todos os times que participaram de 2003 a 2020; (f) Soma da métrica *da* dividida pelo número de times que participaram naquela edição, considerando os anos de 2003 a 2020.



Fonte: Autores (2022)

Figura 2 – Comparação entre as métricas h_0 (pontos com contorno e preenchimento verdes) e *d* (pontos com contorno em preto) para medir o efeito da casa para cada participação de cada time ao longo das edições do Campeonato Brasileiro de Futebol (Série A) de 2003 à 2020. Dos pontos com contorno em preto (métrica *d*), os preenchidos em preto indicam vantagem da casa, em vermelho desvantagem da casa e em cinza indicam ausência de efeito significativo da casa.



Fontes: Autores (2022)

3.2 Inferência para participação de um time em uma única edição da competição e representação longitudinal

A principal novidade do presente estudo é a possibilidade de obtenção de uma informação sobre a vantagem da casa para uma única participação de um time em uma competição. Como um exemplo dessa aplicação da métrica e do respectivo teste, considerou-se o time do Náutico em 2013, que fez 12 pontos em casa e 8 pontos como visitante. A diferença de pontos absoluta, $da = P_c - P_f = 12 - 8 = 4$, e a diferença de pontos relativa, $d = (P_c - P_f) \times 100/(m \times v) = (12-8)\times 100/(19\times 3) = 7{,}02\%$. E a partir da distribuição de todos os valores de *d* obtidos no processo de reamostragem quando os pontos ganhos foram sorteados entre casa e fora (Figura 1c), o intervalo de predição deixando 20% das observações nas caudas foi $IP_{LI} = -14{,}035\%$ e $IP_{LS} = 14{,}035\%$. Como o valor de *d* para o Náutico em 2013 foi de 7,02% e está situado dentro do intervalo de predição foi de -14,0% e 14,0%, então conclui-se que não há evidências suficientes para afirmar que existiu efeito da casa para o Náutico em 2013. Já utilizando a métrica *h*, Náutico obteve $h = (P_c \times 100)/(P_c + P_f) = (12)/(12 + 8) = 60{,}0\%$, e $h_0 = 2h - 100 = 120{,}0 - 100{,}0 = 20{,}0\%$. Embora que o valor de 20% da métrica h_0 indicaria a existência de vantagem da casa, não é possível afirmar que exista efeito positivo da casa a partir do teste sobre a métrica *d*. A diferença encontrada não foi suficiente para afirmar que existe vantagem da casa para o Náutico. Sendo que o valor de $h_0 = 20\%$ do presente estudo, pode ser considerado inflacionado para um time que teve apenas uma vitória e um empate a mais em casa do que como visitante.

Quando a métrica e o teste da métrica do exemplo acima são aplicados à todas 370 participações geradas pelos 43 times nas 18 temporadas do Campeonato Brasileiro de Futebol, obtiveram-se 370 valores de efeito da casa. Desses, uma única participação apresentou efeito da casa negativo (desvantagem da casa; que foi o time do Vitória em 2017), 259 apresentaram efeito da casa positivo (vantagem da casa) e 110 não apresentaram efeito da casa a um nível de significância de 20% (Figura 2). O ano com

mais times com efeito da casa positivo foi 2018 que todos os 20 times participantes apresentaram HA e o ano com menos times com HA foi 2017 com 6 times de um total de 20 participantes (Figura 1d). Já em 2020, que teve todas as partidas durante a pandemia do coronavírus, 10 times tiveram efeito positivo da casa. Seria esperado que na edição de 2020 haveriam os menores valores de vantagem da casa, porém 2017 foi o ano com menores valores. Ainda, poderia ser esperado uma redução na vantagem da casa ao longo do tempo, como discutido em Leite (2017). Porém, os resultados do presente estudo, tanto sobre o número de times com efeito da casa positivo (Figura 1d), quanto a soma da métrica da (Figura 1e) e a soma da métrica *da* dividida pelo número de times (Figura 1f), mostraram expressiva variação da vantagem da casa ao longo do tempo, sem um padrão de queda quando analisado graficamente.

O efeito da casa se mostrou variável entre os clubes (Figura 2), informação coerente com o apontado pela literatura para o Campeonato Brasileiro (Pollard *et al*., 2008) e outras competições de outros países (CLARKE; NORMAN, 1995; GOUMAS, 2017). Ainda, existiu uma expressiva variação para um mesmo clube ao longo do tempo, também verificado em outro estudo que utilizou saldo de gols para acessar a vantagem da casa ao invés de pontos no campeonato inglês (CLARKE; NORMAN, 1995). Sendo que esta visualização ao longo do tempo é importante e foi proporcionada pela métrica e pelo teste apresentado no presente estudo. Por exemplo, o Athletico-PR passou a utilizar gramado sintético na temporada de 2016, ano que apresentou o maior valor observado de efeito positivo da casa para o time. Sendo que em um estudo realizado em país europeu, o gramado sintético proporcionou maior vantagem da casa (van Ours, 2019) e o gramado sintético pode ser uma das causas para esse valor observado para o Athletico-PR em 2016, fato que merece estudos adicionais. Como outro exemplo, pode ser citado o time do Internacional que em 2013 não utilizou o seu estádio e não apresentou vantagem da casa, mas apresentou vantagem da casa em 2014 e 2015 (Figura 2), após a reforma do estádio. Essa detecção de variação ao longo do tempo é um passo importante para auxiliar na explicação dos principais fatores que afetam a

vantagem da casa, questão que ainda não está completamente elucidada na literatura. Seria importante que fosse observada essa variação ano a ano para buscar associações do que pode explicar o efeito positivo da casa.

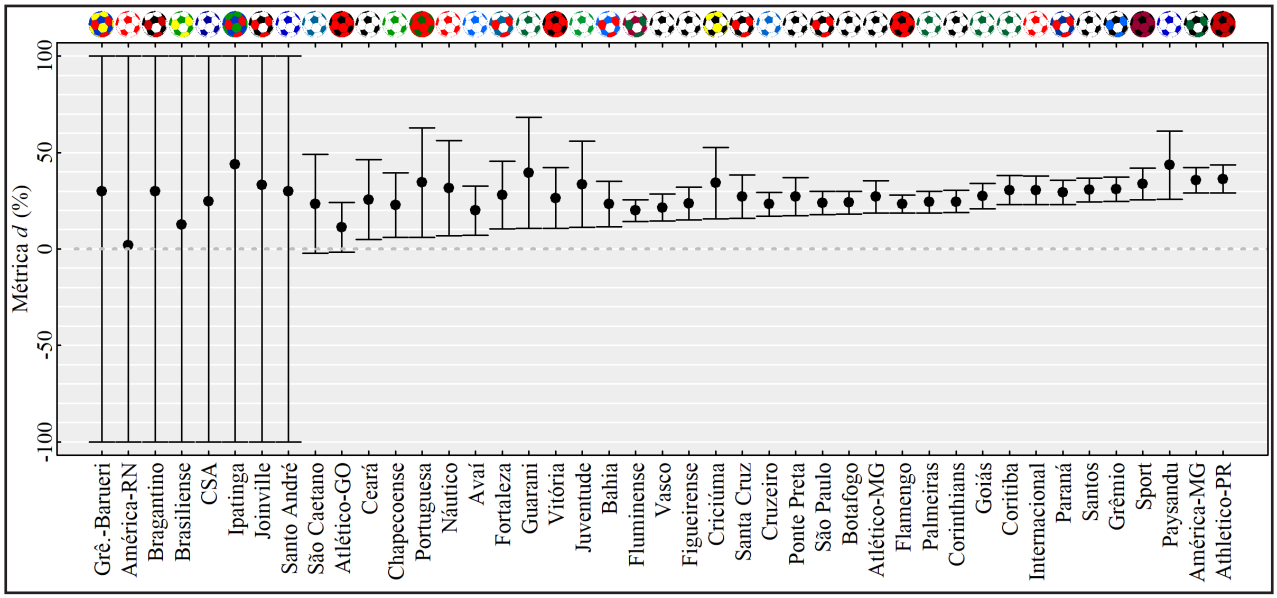
Outro aspecto que pode ser observado é sobre o rebaixamento de times da Série A para a Série B do campeonato. Dos 43 times que participaram no Campeonato Brasileiro Série A, 38 deles foram rebaixados em algum momento. Na Figura 2, é possível visualizar quando um time caiu da Série A para a Série B, que foi quando um ponto não foi sucedido por uma linha. Ainda, em algumas destas participações que terminaram com queda para outra divisão, há efeito da casa positivo significativo no ano anterior, mas no ano da queda não há efeito da casa significativo. Esse padrão de não haver significância no ano que caiu, aconteceu em 14 times, como exemplo o Atlético Goianense em 2007, o Atlético-MG em 2005, Avaí em 2011, entre outros. Eis duas prováveis explicações: (i) se o time está indo mal na tabela, a cobrança da torcida pode passar a ter um efeito negativo sobre o desepenho do time ou; (ii) se um time cai para a outra divisão é porque fez poucos pontos, e com poucos pontos é mais fácil acontecer a situação da informação ser insuficiente para o teste acusar que exista um efeito da casa. Questão essa que merece estudos adicionais.

3.3 Intervalo de confiança para a média populacional de *d*

Quando não é mais realizada uma avaliação de cada participação de um time, mas se obtém uma média amostral $\overline{D}$ da métrica *d* de um número *n* de participações (Tabela 2), pode ser realizado o ranqueamento dos times de acordo com a vantagem da casa (Figura 3). Assim, ranqueando-se pelo intervalo de confiança inferior, os times dos quais se pode esperar maiores valores de efeito positivo da casa são: Athletico-PR, América-MG e Paysandu (Figura 3; Tabela 2). Os três times que se espera um efeito menos expressivo são: São Caetano, Atlético-GO e Ceará. No estudo de Pollard *et al*. (2008) os times com maior vantagem da casa foram o Paysandu, Athletico-PR e Juventude, semelhante ao encontrado no presente estudo. O Paysandu obteve 24,9% no estudo de Pollard *et al*.

(2008), e no presente estudo foi 21,8%. Ainda, os 8 primeiros times expostos na Figura 3 têm o intervalo de confiança em toda a faixa possível da métrica *d*, uma vez que 7 deles participaram uma única vez e isso faz com que não seja possível a obtenção de um intervalo de confiança.

Figura 3 – Intervalo de confiança para o efeito da casa medido pela métrica *d* para todos os times que participaram do campeonato Brasileiro Série A de 2003 a 2020. Em que o ponto do centro é a média da diferença de pontos relativa e as barras verticais nas extremidades representam os limites do intervalo de confiança para a média populacional com 95% de probabilidade de conter um novo valor de vantagem da casa em uma nova participação de um time



Fontes: Autores (2022)

Tabela 2 – Número de participações de cada time (*n*), média amostral da métrica *da* ou diferença de pontos absoluta média ($\overline{DA}$), desvio padrão da métrica *da* (S_{da}), intervalo de confiança para a média populacional de *da* (*IC[ξ;95%]*), média amostral da métrica *d* ou diferença de pontos relativa média ($\overline{D}$), desvio padrão da métrica *d* (S_d), intervalo de confiança para a média populacional de *d* (*IC[δ;95%]*), valores obtidos pela métrica *h* e número de vezes que um time apresentou efeito positivo da casa (n_+). São dados de vantagem da casa para todas as participações dos 43 clubes nas 18 edições do Campeonato Brasileiro de Futebol Série A

Time	n	$\overline{DA}$	S_{da}	$IC[\xi;95\%]$	$\overline{D}$	S_d	$IC[\delta;95\%]$	$\overline{H}$	n_+
América-MG	3	20,3	1,5	16,5; 24,1	35,7	2,7	29,0; 42,3	79,6	3
América-RN	1	1,0			1,8			52,9	0
Athlético-PR	17	21,4	8,4	17,0; 25,7	36,3	14,1	29,0; 43,5	69,3	15
Atlético-GO	5	6,4	5,9	-0,9; 13,7	11,2	10,3	-1,6; 24,1	57,6	2
Atlético-MG	17	15,8	9,3	11,0; 20,5	27,1	16,2	18,7; 35,4	63,8	12
Avaí	6	11,3	6,9	4,0; 18,6	19,9	12,2	7,1; 32,7	64,0	3
Bahia	9	13,9	9,9	6,3; 21,5	23,3	15,3	11,6; 35,0	65,0	4
Botafogo	16	14,0	6,5	10,5; 17,5	24,0	10,9	18,2; 29,8	64,0	10
Bragantino	1	17,0			29,8			66,0	1
Brasiliense	1	8,0			12,7			59,5	0
Ceará	5	14,6	9,6	2,7; 26,5	25,6	16,8	4,8; 46,4	67,2	4
Chapecoense	6	13,0	9,1	3,4; 22,6	22,8	16,0	6,0; 39,6	64,3	3
Corinthians	17	14,3	6,3	11,0; 17,6	24,5	11,2	18,8; 30,3	62,3	10
Coritiba	13	18,0	7,0	13,7; 22,3	30,5	12,6	23,0; 38,1	68,0	10
Criciúma	4	22,0	9,1	7,5; 36,5	34,2	11,6	15,7; 52,7	73,8	3
Cruzeiro	17	13,6	7,1	10,0; 17,3	23,2	12,1	17,0; 29,4	61,9	12
CSA	1	14,0			24,6			71,9	1
Figueirense	11	14,4	8,3	8,8; 19,9	23,6	12,6	15,1; 32,0	64,9	6
Flamengo	18	13,8	6,0	10,8; 16,8	23,3	9,3	18,7; 27,9	61,7	13
Fluminense	18	11,8	7,4	8,2; 15,5	19,9	11,5	14,2; 25,6	61,4	9
Fortaleza	5	17,2	9,1	5,9; 28,5	27,9	14,1	10,4; 45,5	67,6	4
Goiás	13	16,5	7,4	12,0; 21,0	27,5	11,0	20,9; 34,1	65,6	9
Grêmio	17	18,1	6,9	14,6; 21,7	31,0	12,1	24,8; 37,3	65,7	15
Grê.-Barueri	2	17,0	11,3	-84,6; 118,6	29,8	19,8	-148,5; 208,2	70,0	1
Guarani	3	25,7	8,3	5,0; 46,4	39,5	11,6	10,7; 68,4	76,4	3
Internacional	17	17,9	8,5	13,6; 22,3	30,4	14,2	23,1; 37,7	65,0	13
Ipatinga	1	25,0			43,9			85,7	1
Joinville	1	19,0			33,3			80,6	1
Juventude	5	20,6	10,6	7,4; 33,8	33,5	18,1	11,1; 56,0	71,0	4
Naútico	5	18,0	11,3	3,9; 32,1	31,6	19,9	6,9; 56,3	71,1	4
Palmeiras	16	14,0	6,1	10,8; 17,2	24,3	10,5	18,7; 29,9	62,3	11
Paraná	6	18,3	4,7	13,4; 23,3	29,4	6,0	23,1; 35,7	69,6	6
Paysandu	3	29,3	6,0	14,4; 44,3	43,6	7,1	25,9; 61,2	78,5	3
Ponte Preta	9	16,1	7,1	10,7; 21,6	27,2	12,9	17,3; 37,1	67,1	5
Portuguesa	3	19,7	6,5	3,5; 35,8	34,5	11,4	6,1; 62,9	73,2	3
Santa Cruz	2	15,5	0,7	9,1; 21,9	27,2	1,2	16,0; 38,3	76,4	2
Santo André	1	17,0			29,8			70,7	1
Santos	18	17,8	7,1	14,3; 21,4	30,6	12,6	24,3; 36,8	64,9	16
São Caetano	4	15,5	11,4	-2,6; 33,6	23,4	16,1	-2,2; 49,1	62,8	3
São Paulo	18	14,1	7,3	10,4; 17,7	23,9	11,9	17,9; 29,8	60,9	10
Sport	10	19,2	6,6	14,5; 23,9	33,7	11,5	25,4; 41,9	70,7	8
Vasco	15	12,9	7,9	8,5; 17,2	21,5	12,7	14,5; 28,6	62,9	8
Vitória	10	16,0	13,4	6,4; 25,6	26,4	21,9	10,7; 42,1	66,6	7

4 CONSIDERAÇÕES FINAIS

Atingindo o objetivo principal, este trabalho propôs uma métrica, isto é, uma função de variáveis aleatórias para se obter o efeito da casa considerando alguns aspectos. Sendo que o principal aspecto da métrica é o não inflacionamento quando um time ganha poucos pontos, o que atendeu à primeira consideração proposta nos objetivos. Como característica, a métrica exprime o resultado como uma percentagem em relação ao total de pontos conquistados possíveis de serem conquistados, o que permite comparar diferentes competições de pontos corridos ou até diferentes esportes que atribuem pontuação por vitória e por empates (quando existirem empates).

Para satisfazer a segunda consideração proposta nos objetivos, isto é, avaliar se existe evidências para concluir se existe efeito de casa, este trabalho também utilizou um teste, que na forma como foi aplicado levou em consideração se os pontos foram obtidos em empates ou vitórias. Sendo que o teste permite verificar se é possível afirmar que há efeito da casa positivo (vantagem da casa), efeito da casa negativo da casa (desvantagem da casa) ou se não há evidências suficientes para afirmar que exista efeito da casa. Um resultado principal é que existiu o efeito positivo da casa (vantagem da casa) em 73% das participações dos times na competição, sendo que a desvantagem da casa aconteceu apenas uma única vez. Ainda, a métrica e o teste permitiram observar que há expressiva variação na vantagem da casa de um time ao longo do tempo.

Podem ser destacadas algumas limitações da métrica *d*. A primeira delas é que a métrica *d* é relativamente simples de ser obtida, porém, a inferência para uma participação de um time em uma competição exige o uso de recursos computacionais, o que torna mais onerosa a sua obtenção. Outro ponto que pode ser destacado é a sua aplicabilidade para profissionais da área, que ainda não foi avaliada. Ainda, a obtenção da vantagem da casa conjunta para todos os times que participaram de uma competição pode ser obtida utilizando-se a média aritmética dos valores de *d* de cada

time ou pode ser obtido de forma exata, como explicado no apêndice. Sendo que na forma exata a sua obtenção é mais onerosa do que quando comparada à métrica *h*. Por outro lado, para análises com intuito de investigar o que está associado com a vantagem da casa, a métrica *d* possui um comportamento mais estável e que permite obter a informação por ano, sendo que desta forma, poderá contribuir com estudos que tenham esse objetivo.

Estudos futuros poderiam obter valores de vantagem da casa para cada time e com um maior nível de detalhamento, como filtrar os ruídos promovidos pelas perdas judiciais de mando de campo, vendas de mando de campo, entre outros. Ou ainda, estabelecer metodologias para mitigar o efeito desses ruídos na obtenção do valor da vantagem da casa por time. Também seria interessante responder qual métrica consegue predizer melhor valores futuros de vantagem da casa. E ainda, poderia ser investigado quais fatores afetam a vantagem da casa utilizando a presente métrica *d*, que não é inflacionada como a métrica *h*, especialmente no sentido de investigar o que pode estar associado a variação observada para cada clube em cada participação na competição.

AGRADECIMENTOS

À Professora Dra. Josiane Magalhães Teixeira e aos Professores Dr. Person Pereira Neves, Dr. Luiz Alberto Beijo e Dr. Denismar Alves Nogueira pelas contribuições durante o desenvolvimento do artigo. Aos revisores que contribuíram para melhoria do presente artigo. À CAPES, sendo que o presente trabalho foi realizado com o apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

REFERÊNCIAS

CLARKE, S. R.; NORMAN, J. M. Home ground advantage of individual clubs in English soccer. **Journal of the Royal Statistical Society: Series D (The Statistician)**, [*s.l.*], v. 44, n. 4, p. 509–521, 1995.

COURNEYA, K. S.; CARRON, A. V. The Home Advantage in Sport Competitions: A Literature Review. **Journal of Sport and Exercise Psychology**, [*s.l.*], v. 14, n. 1, p. 13–27, 1992.

DAWSON, P.; MASSEY, P.; DOWNWARD, P. Television match officials, referees, and home advantage: Evidence from the European Rugby Cup. **Sport Management Review**, [*s.l.*], v. 23, n. 3, p. 443–454, 2020.

FAJARDO, L.; WERNECK, F.; COELHO, E. F.; MATTA, M. O. A vantagem de jogar em casa em relação às séries do Campeonato Brasileiro de futebol. **Revista Brasileira de Futebol (The Brazilian Journal of Soccer Science)**, [*s.l.*], v. 10, n. 2, p. 25–34, 2019.

GOUMAS, C. Modelling home advantage for individual teams in UEFA Champions League football**. Journal of sport and health science**, [*s.l.*], v. 6, n. 3, p. 321–326, 2017.

JACKLIN, P. B. Temporal changes in home advantage in English football since the Second World War: What explains improved away performance? **Journal of sports Sciences**, [*s.l.*], v. 23, n. 7, p. 669–679, 2005.

LEITE, W. S. S. Home Advantage: Comparison between the Major European Football Leagues. **Athens Journal of Sports**, v. 4, n. 1, p. 65–74, 2017.

MAREK, P., VÁVRA, F. Comparison of Home Advantage in European Football Leagues. **Risks**, [*s.l.*], v. 8, n. 3, p. 87, 2020.

MATOS, R. M.; AMARO, N.; POLLARD, R. How best to quantify home advantage in team sports: an investigation involving male senior handball leagues in Portugal and Spain. **RICYDE Revista Internacional de Ciencias del Deporte**, [*s.l.*], p. 16, v. 59, p. 12–23, 2020.

OLIVEIRA, P. V. S. R.; SILVA, O. W.; ASSIS, J. V.; LAVORATO, V. N.; MOTA-JÚNIOR, R. J.; SILVA, D. C. Vantagem de Jogar em Casa na Série A do Campeonato Brasileiro e na Copa do Brasil. **Revista Brasileira de Futsal e Futebol**, [*s.l.*], v. 12, p. 48, p. 180–186, 2020.

VAN OURS, J. C. A Note on Artificial Pitches and Home Advantage in Dutch Professional Football. **De Economist**, [*s.l.*], v. 167, p. 1, p. 89–103, 2019.

POLLARD, R.; GÓMEZ, M. A. Home advantage analysis in different basketball leagues according to team ability**. Iberian Congress on Basketball Research**, [*s.l.*], v. 4, p. 61–64, 2007.

POLLARD, R.; POLLARD, G. Long-term trends in home advantage in professional team sports in North America and England (1876-2003). **Journal of Sports Sciences**, [*s.l.*], v. 23, n. 4, p. 337–50, 2005.

POLLARD, R.; SILVA, C. D.; MEDEIROS, N. C. Home advantage in football in Brazil: Differences between teams and the effects of distance traveled. **Brazilian Journal of Soccer Science**, [*s.l.*], v. 1, n. 1, p. 3–10, 2008.

R CORE TEAM R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2023. Disponível em: https://www.R-project.org/ . Acesso em: 10 jan. 2023.

RAMOS, L. F. P.; FERNANDES, H. C.; BATISTA, B. D. O. Modelagem matemática para previsão de jogos de futebol. **Revista Sergipana de Matemática e Educação Matemática**, [*s.l.*], v. 6, n. 1, p. 46–64, 2021.

SILVA, C. D.; ABAD, C. C. C.; MACEDO, P. A. P.; FORTES, G. O. I.; NASCIMENTO, W. W. G. Competitive balance in football: A comparative study between Brazil and the main European leagues (2003-2016). **Journal of Physical Education**, v. 29, e2945, 2018.

TILP, M.; THALLER, S. Covid-19 has turned home-advantage into home-disadvantage in the German Soccer Bundesliga. **Frontiers in Sports and Active Living**, [*s.l.*], v. 2, p. 165, 2020.

APÊNDICE A

Cabe ressaltar que a métrica *d* foi construída para ser aplicada na participação de um time em uma edição de uma competição de pontos corridos, porém, a métrica *d* pode ser aplicada a todos os times de uma edição do campeonato ou a todos os times em várias edições do campeonato. Na construção da métrica, o numerador consiste na diferença observada de pontos entre casa e fora e, por sua vez, o denominador consiste na diferença máxima possível de pontos entre casa e fora. Em uma participação de um time em um campeonato, *d* pode ser obtido como:

$$d = \frac{P_c - P_f}{m \times v} \times 100,$$

em que v são os pontos por vitória; P_c são os pontos conquistados em casa; P_f são os pontos conquistados fora de casa e; m é o número de partidas ou rodadas que um time joga em casa em uma edição do campeonato.

No artigo, quando procurou-se obter um valor de *d* para todo o Campeonato Brasileiro, foi utilizada uma média aritmética a partir de todos os 370 valores de *d* obtidos

nas 370 participações dos times nas 18 edições do campeonato. Alternativamente, é possível obter um d_{global} exato de duas maneiras. A primeira maneira é utilizando o número total de partidas (*z*) consideradas, assim, d_{global} pode ser escrito como:

$$d_{global} = \frac{\sum_{i=1}^{z}(p_{c_i} - p_{f_i})}{z \times v} \times 100,$$

em que *i* = 1, 2, ..., *z* é um índice para as partidas que estão sendo consideradas, sendo que *z* é o número total de partidas que existiram; p_{ci} é o número de pontos que o time mandante conquistou por partida; p_{fi} é o número de pontos que o time vistante conquistou por partida e; *v* é a pontuação por vitória.

Uma segunda alternativa é expandindo a definição do *d* para todos os times que participaram de uma edição do campeonato.

Que, ao invés da diferença de pontos de um time, utiliza-se no numerador a soma das diferenças de todos os times. Já no denominador, utiliza-se então a diferença de pontos máxima possível, isto é, multiplicam-se os termos pelo número de times *t*. Ou seja, $d_{campeoanto}$ pode ser obtido como:

$$d_{campeonato} = \frac{\sum_{j=1}^{t}(P_{c_j} - P_{f_j})}{m \times v \times t} \times 100,$$

em que *v* são os pontos por vitória; P_{cj} é o número total de pontos que um time *j* conquistou como mandante; P_{fj} é o número total de pontos que um time *j* conquistou como visitante; *m* é o número de partidas ou rodadas que um time joga em casa em uma edição do campeonato e *j* é um índice que *j* = 1 refere-se ao primeiro time indo de *j* = 1, 2, ..., t, em que *t* é também o número total de times.

Esse $d_{campeonato}$ pode então ser expandido para toda a competição. Introduz-se a participação, isto é, cada vez que um time participa em uma edição do campeonato gera uma participação (*p*). E ainda, destaca-se a inclusão de número de edições (*e*) no

denominador. Ou seja, d_{global} pode ser obtido como:

$$d_{global} = \frac{\sum_{j=1}^{p} (P_{c_j} - P_{f_j})}{m \times v \times t \times e} \times 100,$$

em que *v* são os pontos por vitória; P_{cj} é o número total de pontos que um time conquistou como mandante em uma participação *j;* P_{fj} é o número total de pontos que um time conquistou como visitante em uma participação *j*; *m* é o número de partidas ou rodadas que um time joga em casa em uma edição do campeonato; *j* é um índice que *j* = 1 refere-se à primeira participação indo de *j* = 1, 2, ..., *p*, em que *p* é também o número total de participações e; *e* é o número de edições.

No caso do cálculo da métrica *d* para todas as edições do Campeonato Brasileiro de Futebol de 2003 a 2020, há que ser considerado que em 2003 e 2004 havia 24 times e eram 46 rodadas, sendo 23 como mandante. Já em 2005 foram 42 rodadas de 22 times. De 2006 a 2020, o número de rodadas foi sempre 38 e 20 foi o número de times. Portanto, o denominador não pode apenas multiplicar pelo número de edições, mas foi necessário abri-lo para incorporar essa diferença no número de times, como segue,

$$\begin{aligned} d_{global} &= \frac{\sum_{j=1}^{p} (P_{c_j} - P_{f_j})}{m \times v \times t \times e} \times 100 \\ &= \frac{12879 - 7014}{19 \times 3 \times 20 \times 15 + 21 \times 3 \times 22 \times 1 + 23 \times 3 \times 24 \times 2} \times 100 \\ &= \frac{5865}{21798} \times 100 \\ &= 26{,}91\%, \end{aligned}$$

em que a diferença relativa de pontos global para o Campeonato Brasileiro de 2003 a 2020 foi de 26,91%. Ainda, o valor de 21798 poderia ser subtraído em 3 pontos, pois uma partida de 2016 não existiu e não somou pontos.

APÊNDICE B

Código para utilização no Software R. O código retorna o limite superior e o limite interior do intervalo de predição para uma dada configuração que aconteceu para um time e também retorna a métrica diferença de pontos relativa (*d*) e a diferença de pontos absoluta (*da*) observada para o time em um dado ano. O número recomendado de reamostras é de 1.000.000.

A função teste_reamostragem(·) tem as seguintes entradas: vc é o número de vitórias em casa; ec é o número de empates em casa; dc é o número de derrotas em casa; vv é o número de vitórias como visitante; ev é o número de empates como visitante, dv é o número de derrotas como visitante, n_simu é o número de reamostras obtidas, percentil_bilateral é a proporção (em número decimal) de observações que ficaram nas duas caudas somadas e recomenda-se utilizar 0,20; pto_v são os pontos por vitória e; pto_e são os pontos por empate.

```
#Métrica e teste
teste_reamostragem<-function(vc, ec, dc, vv, ev, dv, n_simu,
percentil_bilateral, pto_v, pto_e)
{
p1<-vc+ec+dc ; p2<-p1+1; p3<-p1*2
conjunto_c<-c(rep(3,vc),rep(1,ec),rep(0,dc))
conjunto_v<-c(rep(3,vv),rep(1,ev),rep(0,dv))
conjunto<-c(conjunto_c,conjunto_v)
reamostras<-matrix(0,n_simu,1)
for(i in 1:n_simu){
cr<-sample(conjunto)
reamostras[i]<-((sum(cr[1:p1])-sum(cr[p2:p3]))*100)/((p3/2)*pto_v)}
ip_linf<-sort(reamostras)[(percentil_bilateral/2)*n_simu]
ip_lsup<-sort(reamostras)[(1-percentil_bilateral/2)*n_simu]
```

```
da<-vc*pto_v+ec*pto_e - vv*pto_v - ev*pto_e
d<- (da*100)/((vc+ec+dc)*pto_v)
return(list(limite_inferior_predicao = ip_linf,
limite_superior_predicao = ip_lsup,
diferenca_pontos_relativa=d, diferenca_pontos_absoluta=da))
}
#Exemplo para o Náutico em 2013
teste_reamostragem(vc=3, ec=3, dc=13, vv=2, ev=2, dv=15, n_simu=1000000,
percentil_bilateral=0.2, pto_v=3, pto_e=1)

#Resultados
#$limite_inferior_predicao
#[1] -14.03509
#
#$limite_superior_predicao
#[1] 14.03509
#
#$diferenca_pontos_relativa
#[1] 7.017544
#
#$diferenca_pontos_absoluta
#[1] 4
```

Contribuição de autoria

1 – Giovani Festa Paludo

Mestrando em Estatística Aplicada e Biometria
https://orcid.org/0000-0002-8046-8409 • gfpaludo@gmail.com
Contribuição: Conceituação, Metodologia, Validação, Investigação, Escrita Primeira Redação, Escrita - Revisão e Edição, Visualização de dados

2 – Nikolas Neves de Figueiredo

Graduando em Licenciatura em Matemática
https://orcid.org/0000-0003-4911-7245 • nikolasfig@gmail.com
Contribuição: Validação, Análise Formal, Escrita - Revisão e Edição, Visualização de Dados

3 – Eric Batista Ferreira

Doutor em Estatística e Experimentação Agropecuária
https://orcid.org/0000-0003-3361-0908 • eric.ferreira@unifal-mg.edu.br
Contribuição: Conceituação, Metodologia, Análise Formal, Investigação, Escrita - Revisão e Edição, Visualização de Dados, Supervisão

Como citar este artigo

PALUDO, G. F.; FIGUEIREDO, N. N.; FERREIRA, E. B. Proposta de uma métrica para o efeito da casa baseada em pontos ganhos. **Ciência e Natura**, Santa Maria, v. 45, e3, 2023. DOI: https://doi.org/10.5902/2179460X68508.