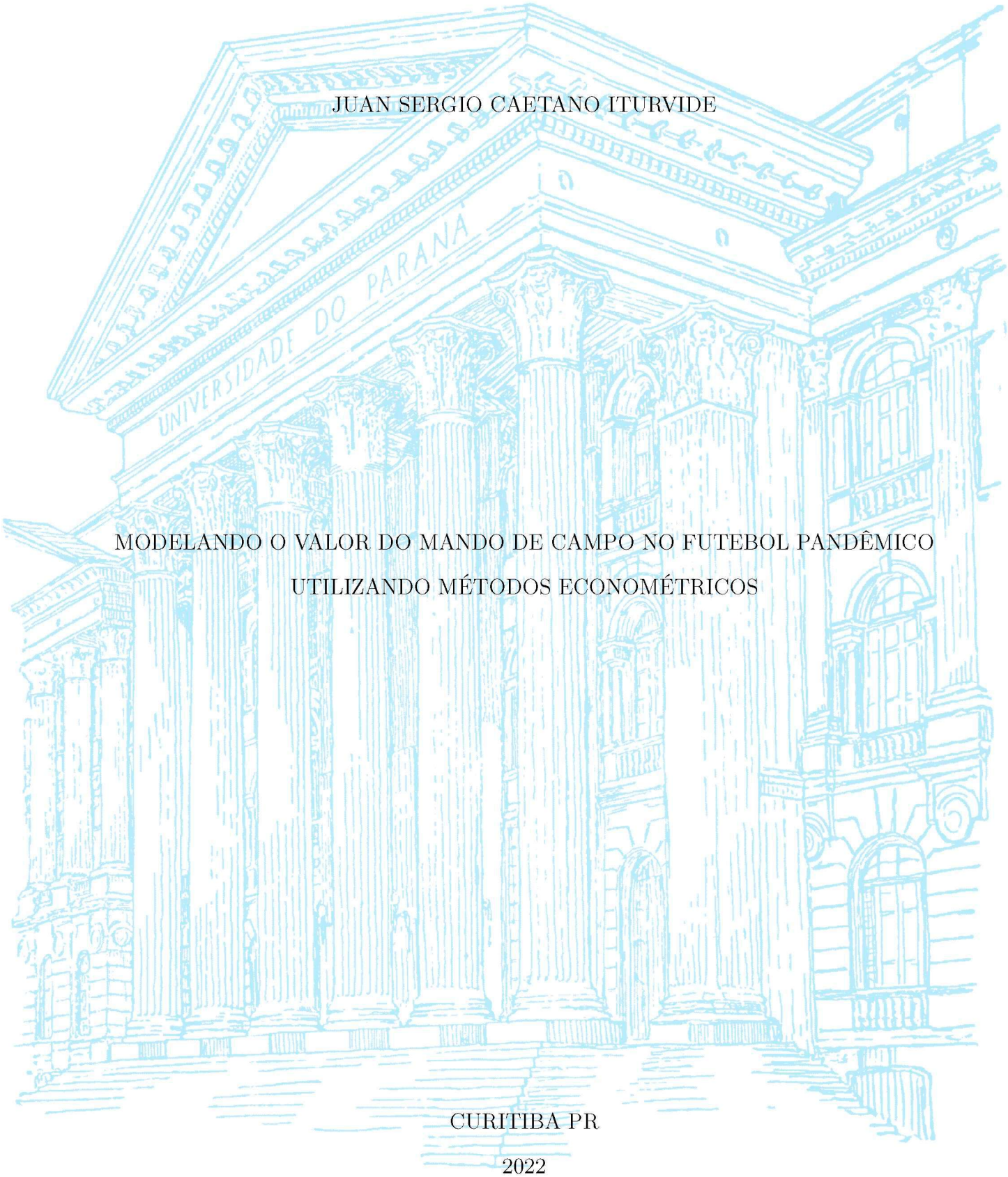
UNIVERSIDADE FEDERAL DO PARANÁ

JUAN SERGIO CAETANO ITURVIDE



MODELANDO O VALOR DO MANDO DE CAMPO NO FUTEBOL PANDÊMICO UTILIZANDO MÉTODOS ECONOMÉTRICOS

CURITIBA PR

2022

JUAN SERGIO CAETANO ITURVIDE

Modelando O Valor Do Mando De Campo No Futebol Pandêmico Utilizando Métodos Econométricos

Monografia apresentada como requisito final à conclusão do Curso de Bacharelado em Ciências Econômicas, Setor de Ciências Sociais Aplicadas, da Universidade Federal do Paraná.

Área de concentração: *Econometria; Microeconomia Aplicada.*

Orientador: Profº. Dr. João Basílio Pereima Neto.

CURITIBA PR

2022

AGRADECIMENTOS

À Deus, pelo honra de viver esse momento.

À minha mãe, pelo carinho, compreensão e amor. À minha irmã, pela companhia e fraternidade. À meu pai, meu maior herói e exemplo, que me ensinou tudo que sei.

À meu avô, que olha por mim todos os dias.

À Beatriz, Priscila, Luis, Lucas, João Vitor, João Pedro, Enrico, Thiago e Nilo por todas as manhãs, trabalhos e aulas. À Ana por compartilhar comigo a paixão pela ciência.

À meus mentores e professores, que contribuíram para minha formação e ensinaram a função social do economista.

Aos amigos que sempre estiveram lá por mim.

RESUMO

Desde o início do Campeonato Brasileiro de 2020, os jogos disputados no decorrer da competição foram realizados sem a presença de público, devido à pandemia do novo coronavírus e os decretos das autoridades locais e federais. Logo, questiona-se como a ausência da torcida mandante (e visitante, em menor nível) tenha influenciado no resultado da partida. O objetivo desta monografia será avaliar o impacto do mando de campo durante o período pandêmico, e como ele afeta o placar final, entendendo assim se a ausência dos fãs nas arquibancadas tenha nivelado o desempenho das duas equipes. Para tal, o objetivo específico do trabalho será criar um modelo econométrico capaz de quantificar e isolar o efeito de uma equipe ser mandante de um determinado jogo. Foi evidenciado que apesar da ausência de torcida, é possível concluir que a equipe mandante ainda tem certa vantagem em relação ao visitante. Porém, esta teve acentuada redução quando comparada aos níveis pré pandemia.

Palavras-chave: Futebol; Vantagem do Mando de Campo; Coronavírus; Econometria

ABSTRACT

Games during the 2020 Brazilian National Football Tournament were played without the presence of fans due to the coronavirus pandemic and federal/local government law enforcements. Therefore, many wondered how the lack of home (and visiting fans, in minor case) would impact match results. The main goal of this work is to analyze home field advantage during the pandemic period, and how it affects the final score, explaining if the lack of supporters is leveling team performance. Therefore, the secondary goal of this work is to create an econometric model capable of quantife and isolate home field advantage during a giving match. Our results show that although home field advantage is still influencing the final score, it's effect faced major reduction when compared to pre pandemic data.

Keywords: Football; Home Field Advantage; Coronavirus;Econometrics

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 Introdução

Desde a virada do século XXI, a análise de dados têm sido cada vez mais utilizados em esportes, principalmente coletivos, ao redor do mundo como uma forma de ganhar vantagem competitiva. Um dos exemplos que mais chamam a atenção e que voltou a ganhar notoriedade recentemente devido à uma adaptação cinematográfica é o relato do livro **Moneyball: The Art Of Winnig An Unfair Game**, de Lewis (2003).

No relato jornalístico, Lewis (2003) conta sobre o dia a dia do time de beisebol americano Oakland Athletics e seu general manager[1] Billy Beane. O time, localizado em uma cidade pequena na Califórnia, foi pioneiro no uso de estatísticas para avaliar e desenvolver jogadores, conseguindo competir em igualdade com gigantes do esporte como Yankees (Nova York) e Red Sox (Boston), mesmo com uma folha salarial muito menor e com uma capacidade reduzida de realizar contratações[2].

As ideias postas em prática por Beane e difundidas mundialmente pelo livro de Lewis não só mudaram o rumo do beisebol moderno, como também de todos os outros grandes esportes. O que era algo extremamente marginalizado na indústria da época se tornou *mainstream*[3]. Hoje em dia muitas equipes investem altos valores em departamentos de análise estatística e de dados (Morgulev et al., 2018), e com o futebol não é diferente.

Um dos casos mais citados atualmente é o do Liverpool Football Club, como aponta Naicker (2021). O time inglês, que tem como grupo proprietário a Fenway Sports Group, mesma do Boston Red Sox, uma das franquias de beisebol que mais abraçaram o uso de dados após a revolução de Beane, usa constantemente modelagens estatísticas para avaliar seus jogadores (Schoenfeld, 2019).

Desde o início do Campeonato Brasileiro de 2020, os jogos disputados no decorrer da competição foram realizados sem a presença de público, devido à pandemia do novo coronavírus e os decretos das autoridades locais e federais. Logo, questiona-se como a ausência da torcida mandante (e visitante, em menor nível) tenha influenciado no resultado da partida.

O popular ditado de que "torcida ganha jogo"[4] será posto à prova no decorrer do trabalho. O objetivo desta monografia será avaliar o impacto do mando de campo durante o período pandêmico, e como ele afeta o resultado da partida, entendendo assim se a ausência dos fãs nas arquibancadas tenha nivelado o desempenho das duas equipes. Para tal, o objetivo específico do trabalho será criar um modelo econométrico capaz de quantificar e isolar o efeito de uma equipe ser mandante de um determinado jogo.

Será utilizado como amostragem, primariamente, uma base de dados contendo jogos da temporada 2020[5], tendo em vista que esta foi a única inteiramente disputada com portões fechados, durante o período mais crítico da pandemia. Depois, serão utilizados dados das temporadas 2018 e 2019, tentando entender se a vantagem do mandante diminuiu ou permaneceu a mesma quando comparadas partidas com e sem torcedores. Como foi citado a área de análise de dados, o modelo utiliza como *inputs*[6] métricas avançadas de

[1] Gerente Geral, responsável pela adiministração do elenco de um time nos esportes americanos.

[2] Segundo o site Spotrac, para a temporada de 2022, a folha salarial de Yankees e Rex Sox será de 250 e 202 milhões de dólares, respectivamente. Oakland fechara a temporada com 48 milhões em folha.

[3] Algo amplamente aceito e utilizado.

[4] Clichê comum na cultura do futebol brasileiro.

[5] Ao contrário de 2021, que teve a volta do público no início do torneio.

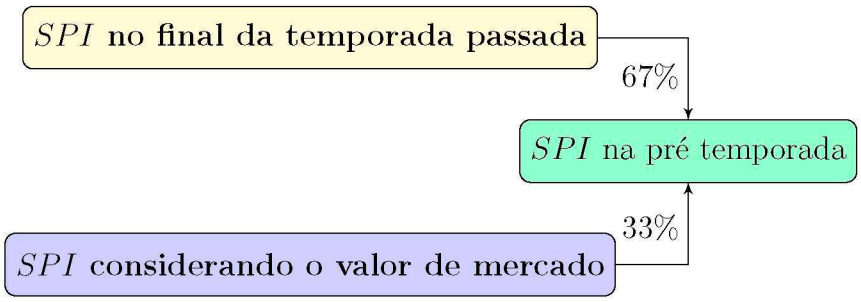
[6] Variáveis explanatórias

avaliação de desempenho, que normalmente não são tão acessíveis ao grande público como estatísticas convencionais de gols e assistências, e que serão explicadas nas seções a seguir. Os números utilizados na montagem da base e as variáveis do modelo foram retirados do site americano de análise estatística FiveThirtyEight.

1.1 *Soccer Power Index* (SPI)

A principal variável que será utilizada no modelo para quantificar a qualidade dos times será o SPI, ou *Soccer Power Index*. Desenvolvido originalmente pela ESPN, o maior rede de jornalismo esportivo do mundo, o SPI é subdividido entre duas outras métricas: O SPI ofensivo, criado para avaliar o ataque, representando o número de gols que uma determinado time marcaria contra uma hipotética equipe mediana em um campo neutro, e o SPI defensivo, que representa o número de gols esperados que o time sofra na mesma situação (Boice, 2020). Conjuntamente, elas formam o SPI Global do time.

Para facilitar o desenvolvimento desta pesquisa e o funcionamento do próprio modelo, utilizaremos o SPI Global. No início de cada temporada, o SPI é definido primariamente pela avaliação do time ao final do torneio anterior e o valor de mercado de seus jogadores após o período de transferências, utilizando o algoritmo do site Transfermarkt, que determina uma cifra monetária (em Euros) para cada atleta do elenco baseado em seu desempenho, como é visto no esquema 1.1.



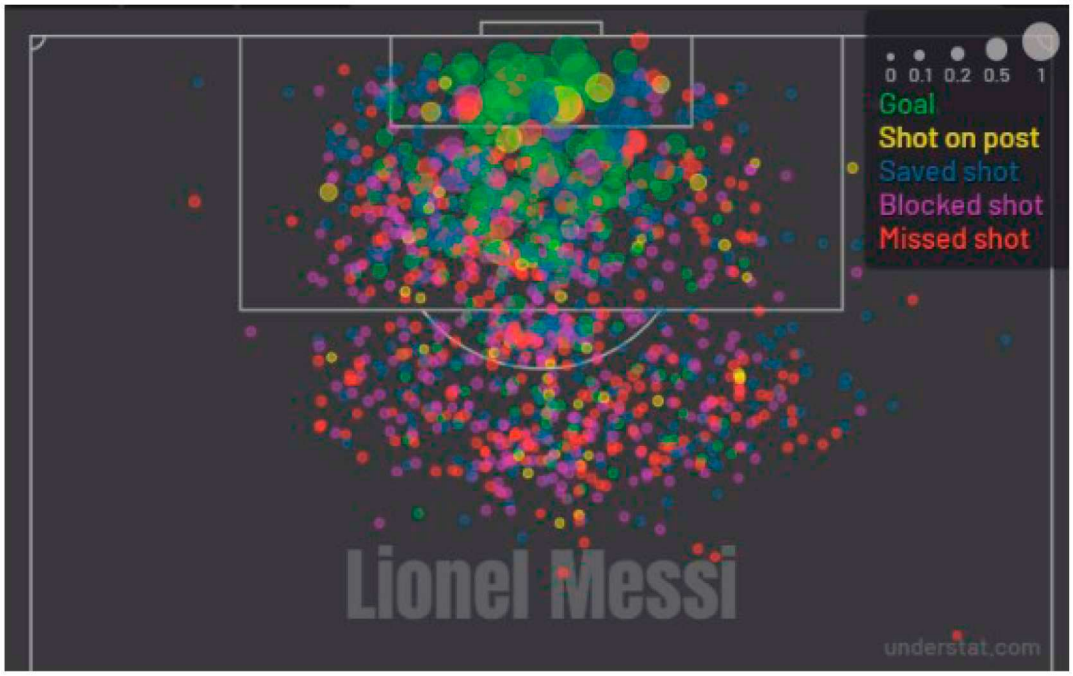
O valor da métrica, então, muda a cada partida levando em consideração o desempenho durante o jogo e o SPI Global do próprio adversário. Essas alterações realizadas jogo a jogo com base no desempenho levam em consideração três outras métricas: *Shot-Based Expected Goals* (Gols Esperados Baseados nos Chutes), *Non-Shot Expected Goals* (Gols Esperados Baseados em Não-Chutes) e *Adjusted Goals* (Gols Ajustados). O uso dessas estatísticas é justificado por serem muito mais confiáveis do que gols e assistências, dado que o futebol é naturalmente um *low-scoring sport*, ou seja, um número exacerbado de gols é raramente visto nos jogos (Boice, 2020).

É importante ressaltar que as três métricas são base tanto do SPI ofensivo quanto para o defensivo, já que se está mensurando o desempenho do time adversário também.

1.1.1 *Shot-Based Expected Goals* (xG)

Rathke (2017) destaca que, durante uma partida, a posição de onde uma finalização ocorre e a velocidade do chute impactam diretamente na probabilidade de que um gol ocorra. Basicamente, o *Shot-Based Expected Goals* (xG) **é a representação da quantidade de gols que um time deveria ter marcado durante a partida baseado em suas finalizações**. O xG representa a probabilidade de um chute ser bem sucedido levando em consideração o local, o jogador responsável e com qual perna ele finalizou.

Figura 1.1: xG das Finalizações de Lionel Messi (2022)



Fonte: Unsertat (2022).

A figura 1.1 exibe todas as finalizações do atacante argentino Lionel Messi durante a última temporada, com as cores representando o resultado respectivo. O tamanho de cada ponto no gráfico representa o xG, variando em um intervalo entre 0 e 1. É possível notar que finalizações próximas ao gol adversário resultaram em uma maior probabilidade de sucesso, principalmente as que foram feitas utilizando a perna esquerda (dado que o jogador é canhoto) e sem marcação defensiva.

A métrica, amplamente difundida na comunidade de análise de dados futebolística, pode ser construída através de diferentes metodologias e variáveis (Robberechts & Davis, 2020), e tem sido uma das principais formas de analisar o desempenho dos jogadores utilizadas em times profissionais (Eggels, 2016). Neste caso, como citado anteriormente, para o SPI, estaremos baseando nosso estudo no modelo do **FiveThirtyEight** por ser disponível gratuitamente ao público geral e de fácil acesso. Este também será o caso das demais estatísticas apresentadas a seguir.

1.1.2 *Non-Shot Based Expected Goals* ($NSxG$)

Ao contrário do xG, o *Non-Shot Based Expected Goals* ($NSxG$) se baseia em ações de jogo que não sejam finalizações. Em resumo, o $NSxG$ **mensura a quantidade de gols que um time deveria ter marcado baseado nas ações ofensivas que não sejam chutes, ou seja, passes, interceptações e etc** (Boice, 2020).

Supondo que um jogador realize um passe vertical, que mova a bola até um ponto muito mais próximo ao gol adversário, o $NSxG$ é responsável por indicar a probabilidade de gol gerada por essa ação ofensiva isolada.

Seguindo a mesma lógica, passes não progressivos em direção ao gol geram um valor negativo, assim como em situações que a bola seria interceptada por um jogador adversário.

Tabela 1.1: Apresentação das Estatísticas

	Everton	Manchester City
Adjusted Goals	3,5	0,0
Shot-Based xG	0,4	0,7
Non-Shot xG	0,7	2,7

Fonte: Boice (2020).

1.1.3 *Adjusted Goals* (aG)

Ao contrário das últimas duas estatísticas apresentadas, o *Adjusted Goals* (aG) tem como finalidade a **contextualização da partida, ajustando os placares dadas as condições que levaram ao resultado** (Boice, 2020).

Alguns exemplos seriam gols marcados por times com diferença no número de jogadores em relação a equipe adversária (valendo mais para times em inferioridade e menos para times com superioridade numérica, caso haja expulsões durante a partida), e a filtragem do chamado *garbage time*[7], dando menos valor aos gols marcados nos momentos finais, quando um dos times já tem muitos tentos marcados. Isto é, se um clube está ganhando por quatro gols de diferença, um quinto gol teria uma diferença de valor marginal muito menor quando comparado ao segundo ou primeiro.

A tabela 1.1 exemplifica como as métricas funcionam. O Everton derrotou o Manchester City por um placar de 4 a 0, mas o xG e o $NSxG$ revelam que os *citizens*[8] jogaram melhor e criaram ações ofensivas mais interessantes que o Everton, o que não é visto no placar final da partida.

1.2 *Home Field Advantage* (HFA)

O fenômeno descrito como *Home Field Advantage*[9] (ou simplesmente HFA) tem sido objeto de estudo de múltiplas pesquisas, de diferentes áreas científicas como psicologia, economia e estatística (Benz & Lopez, 2021), analisando não só o futebol mas também outras modalidades onde o mandante aparenta ter grande vantagem.

Existem diversas hipóteses de como a HFA acaba se manifestando. Moskowitz & Wertheim (2012) sugerem que esta surge através da influência que a torcida local exerce nos árbitros, seja pelo barulho em si (Unkelbach & D., 2010) ou pela pura pressão (Garicano et al., 2005), os juízes da partida tendem a tomar decisões favoráveis ao time local (Benz & Lopez, 2021).

Uma abordagem empírica adequada seria o contraste de jogos com e sem a presença de fãs nas arquibancadas, onde o senso comum diria que a ausência de torcida resultaria em um decréscimo de HFA (Benz & Lopez, 2021). Estudos como os de Pettersson-Lidbom & Priks (2010), utilizando dados do futebol italiano em 2007, e de Bryson et al. (2020), tendo como base duas décadas de futebol europeu, corroboram com a ideia de diminuição de vantagem do mandante.

Com a pandemia do COVID-19 e o fechamento de estádios e complexos esportivos ao redor do mundo, foi criada uma oportunidade de melhor analisar como a torcida (ou a ausência dela) tem impactado no desempenho das equipes e no resultado das partidas.

[7] Termo utilizado para situações da partida em que ela já esta decidida, como os momentos finais de uma goleada.

[8] Apelido do Manchester City.

[9] Vantagem do Mando de Campo, em tradução literal.

Tabela 1.2: Estudos Prévios

Resultados - Levantamento Bibliográfico			
Estudo	Ligas	Método	Resultados
Sors et al. (2020)	8	Correlação	Diminuição de HFA
Leitner & Richlan (2020)	8	Correlação	Diminuição de HFA
Endrich & Gesche (2020)	2	Regressão Linear	Diminuição de HFA
Fischer & Haucap (2020)	3	Regressão Linear	Incertos
Dilger & Vischer (2020)	1	Regressão Linear, Correlação	Diminuição de HFA
Krawczyk & Strawinski (2020)	4	Regressão Linear	Incertos
Ferraresi & Gucciardi (2020)	5	Regressão Linear	Diminuição de HFA
Reade et al. (2020)	7	Regressão Linear	Diminuição de HFA
Sanchez & Lavin (2020)	8	Regressão Linear, Correlação	Incertos
Scoppa (2020)	10	Regressão Linear	Diminuição de HFA
Cueva (2020)	41	Regressão Linear	Diminuição de HFA
McCarrick et al. (2021)	15	Regressões Linear, Poisson	Diminuição de HFA
Bryson et al. (2020)	17	Regressões Linear, Poisson	Incertos
Benz & Lopez (2021)	17	Rregressão de Poisson	Incertos

Fonte: Benz & Lopez (2021).

Benz & Lopez (2021) compilaram uma série de mais de 10 estudos com este objetivo em comum, tendo como base partidas de futebol de diferentes ligas. Foram utilizadas uma série de métodos estatísticos para quantificar a HFA e se esta teve seu resultado afetado pelo fechamento dos portões.

Os resultados obtidos estão listados na tabela (1.2). Como visto, muitos estudos sugerem uma diminuição do fator mando de campo durante a pandemia. Ao contrário desta monografia, porém, estudos sobre o Campeonato Brasileiro são escaços, sendo majoricamente utilizados dados de ligas europeias. Cueva (2020) inputa dados de jogos do Brasileirão em seu modelo, mas estes são associados aos de outras 40 ligas. Não se tem, portanto, uma noção de com a HFA se comportou exclusivamente no Brasil durante a pandemia.

2 Materiáis e Métodos

Para esta pesquisa, foram selecionados dados do Campeonato Brasileiro de 2020 da data base do **FiveThirtyEight** no Github, um repositório online, criando uma nova tabela onde atribuímos um código ID para cada partida com os times envolvidos, o resultado do jogo e os SPI de cada uma das equipes naquele momento.

Como será explicado nas próximas seções, usaremos a diferença de gols entre os times de determinada partida como variável resposta, já que o resultado de um jogo do Campeonato Brasileiro não é binário, tendo em vista que este poderia terminar empatado, e um ajuste convertendo o placar em pontos conquistados poderia resultar em uma piora do modelo, pois vitórias por cinco e um gols de diferença podem representar uma distinção no desempenho, e isso não seria possível ilustrar com pontos.

Foram criados 377 códigos ID, simbolizando todos os jogos realizados durante os dois turnos do Brasileirão de 2020, com a exceção de três partidas que não serão utilizados pois contem os dados de xG e $NSxG$ faltantes. No exemplo da tabela 2.1, o mandante Fortaleza (Game $ID = 2$) perdeu a partida para o visitante Athletico (0) em dois a zero (-2 de saldo final). A tabela ilustra os SPI de cada time antes do jogo acontecer, bem com as diferenças de xG (xG do Fortaleza - xG do Athletico) e de $NSxG$ entre o mandante e o visitante desempenhadas no decorrer da partida. O intuito é que o modelo possa captar tanto o momento das equipe pré partida quanto seu desempenho durante esta. Ao mesmo tempo, a diferença de SPI na tabela pode revelar qual dos times tinha um momento melhor antes da partida (positivo para o mandante e negativo para o visitante).

2.1 Análise dos Dados

As variáveis utilizadas no modelo, apresentadas na base descrita anteriormente, serão primeiramente analisadas de maneira descritiva, identificando as tendências, variações e correlações entre os dados. Os resultados obtidos serão expostas nas seções subsequentes.

O objetivo será identificar o comportamento das variáveis explanatórias escolhidas antes que elas sejam incluídas no modelo, tornando a análise mais completa e compreensível.

2.1.1 Estatísticas descritivas

Levando em consideração as métricas descritivas, como **média, amplitude dos valores, desvios padrões e variâncias**, podemos apresentar os dados de acordo com

Tabela 2.1: Tabela de Dados

Game ID	Time Mandante	Saldo	Dif. SPI	Dif. xG	Dif. NSxG
1	Coritiba	-1	-14,84	-0,85	-0,48
2	Fortaleza	-2	-11,09	-1,66	0,41
3	Flamengo	-1	22,01	1,05	0,87
⋮	⋮	⋮	⋮	⋮	⋮
377	Vasco Da Gama	1	0,62	0,95	0,15

Fonte: O Autor (2022).

a tabela 2.1, apresentada a seguir, para os dados de Diferença de Gols ($Saldo$), SPI ($Dif.SPI$), xG ($Dif.xG$), $NsxG$ ($Dif.NSxG$):

Tabela 2.2: Estatística Descritiva

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
$Saldo$	377	0.326	1.513	-4	-1	1	5
$Dif.SPI$	377	0.134	12.952	-34.13	-9.67	9.42	31.15
$Dif.xG$	377	0.264	1.108	-2.62	-0.47	0.99	5.38
$Dif.NSxG$	377	0.24	0.881	-2.15	-0.27	0.74	3.1

Fonte: O Autor (2022).

Os dados também foram dispostos em uma matriz de correlação. Utilizando as Diferenças de Gols ($Saldo$), SPI, xG e $NsxG$, conforme figura 2.1.

Figura 2.1: Matriz de Correlação

Fonte: O Autor (2022).

Os resultados também podem ser dispostos por gráficos de distribuições, tornando possível a observação da dispersão dos dados em torno da respectiva média de cada variável.

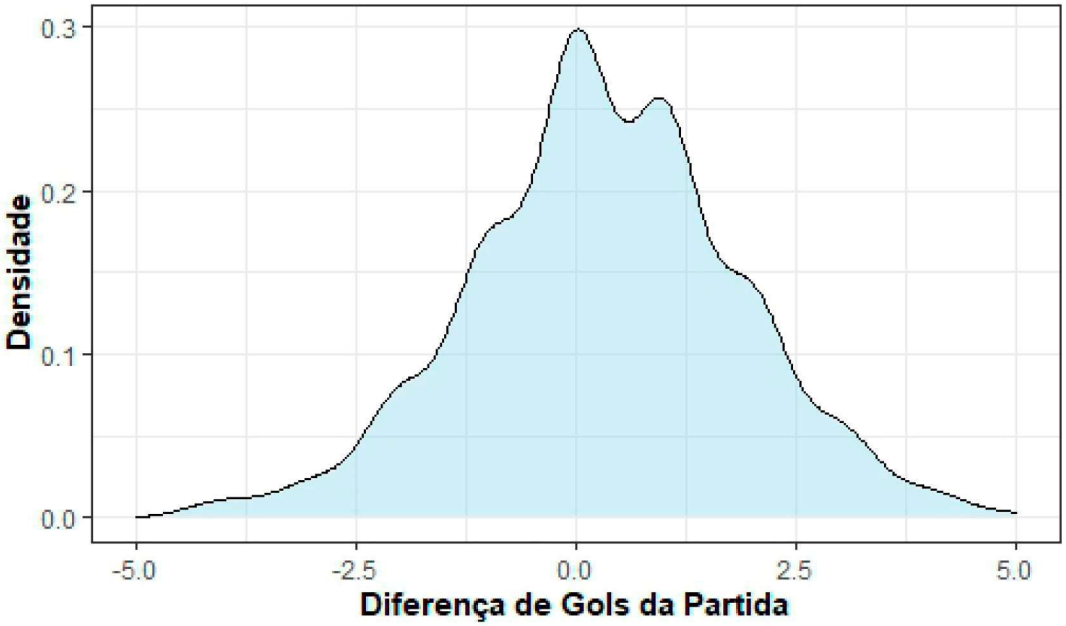
Como podemos ver, a distribuição revela uma certa concentração (pico) de diferença de gols na parte positiva da distribuição para os mandantes e, por consequência, uma na parte negativa para os visitantes. Isso significa que os mandantes tenderam a marcar mais gols que seus contrários.

Foi revelada também uma concentração em certos placares específicos, tendo em vista a dinâmica do futebol como um esporte de *low-scoring* e pouca variedade de resultados prováveis.

A distribuição entre os SPI revelou um concentração muito próxima da média 0, o que diz muito sobre o equilíbrio do campeonato brasileiro quando usada uma métrica capaz de isolar o desempenho real de cada time em um só número.

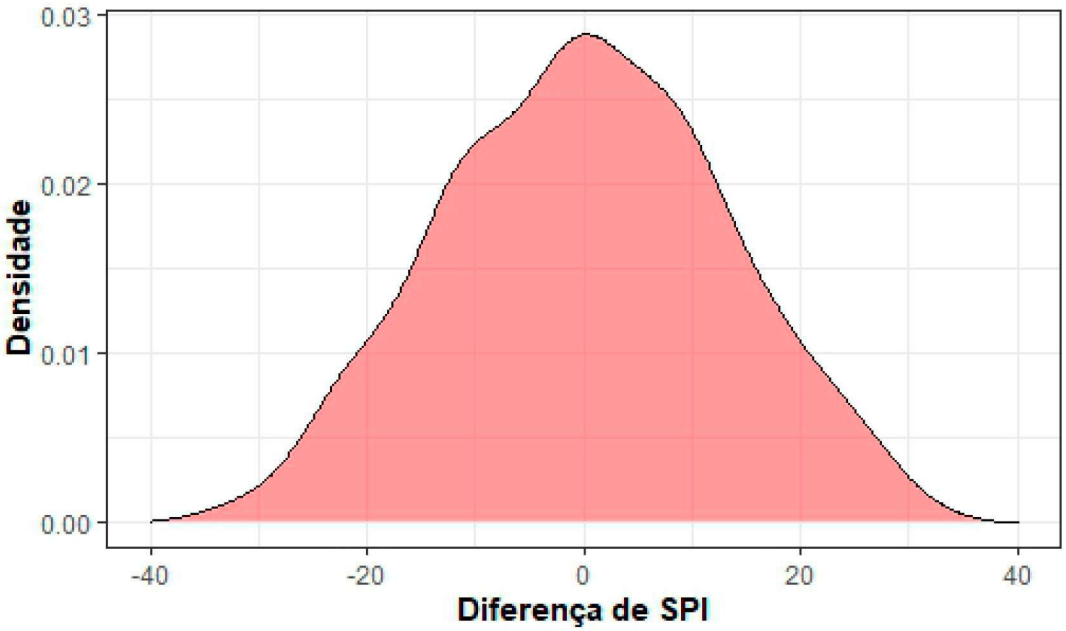
Como evidenciado nas distribuições de xG e $NSxG$, há uma leve tendência de números positivos, o que significa uma vantagem ao time mandante, corroborando com as tabela 2.2, onde as médias estão em torno de 0,264 e 0,24, respectivamente.

Figura 2.2: Distribuição de Diferença de Gols



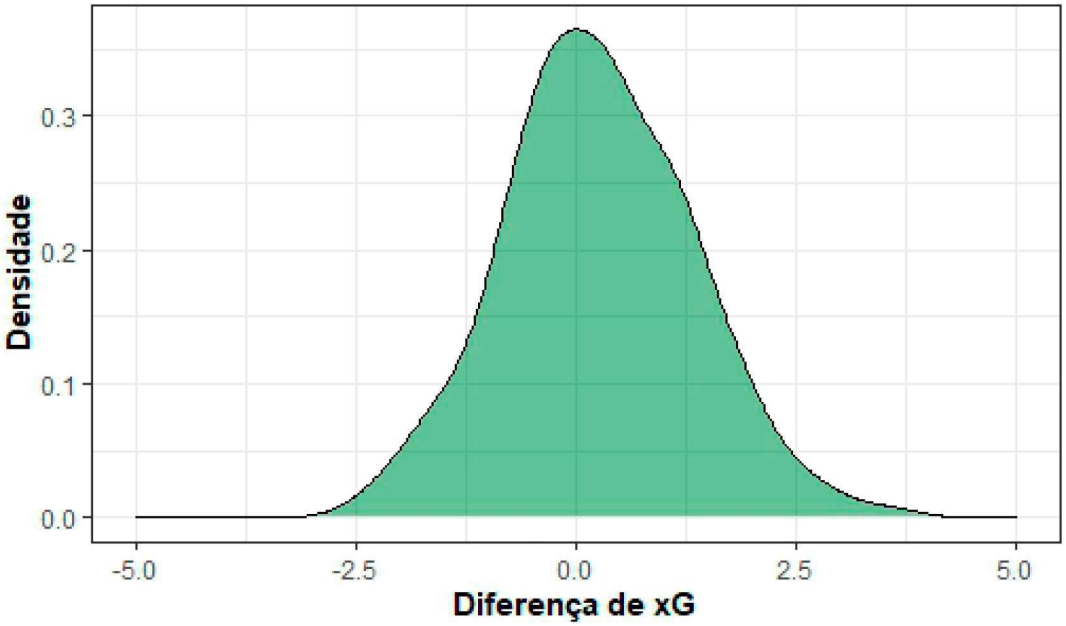
Fonte: O Autor (2022).

Figura 2.3: Distribuição de SPI



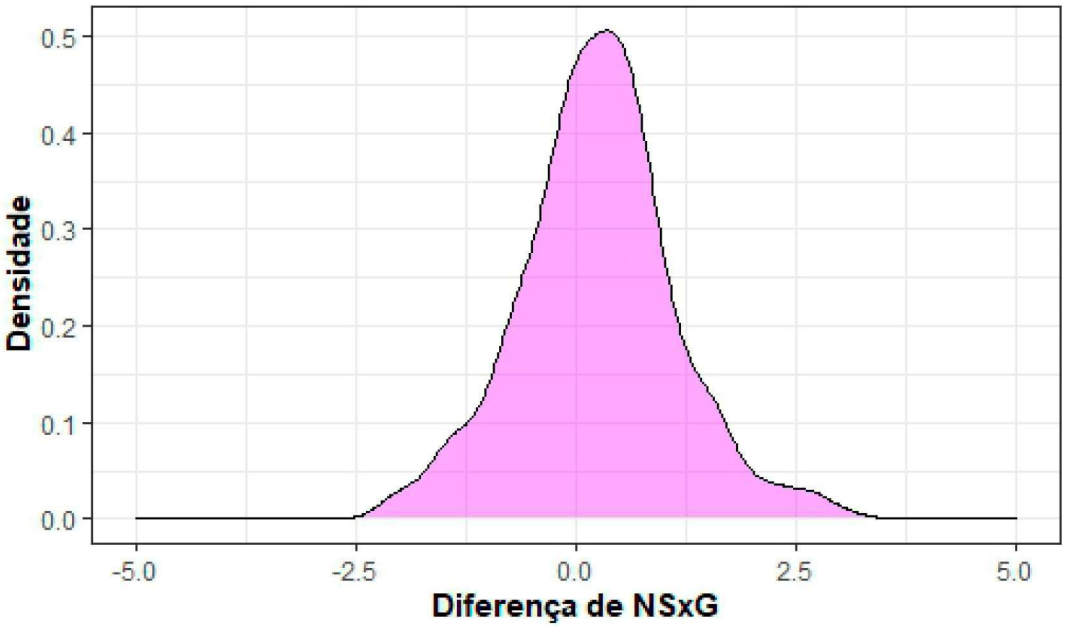
Fonte: O Autor (2022).

Figura 2.4: Distribuição de xG



Fonte: O Autor (2022).

Figura 2.5: Distribuição de NSxG



Fonte: O Autor (2022).

3 Modelagem

Nesta pesquisa, o modelo tentará captar o efeito global do mando de campo no resultado da partida, levando em consideração a qualidade de ambos os times envolvidos antes da partida, através dos SPI pré estabelecidos, e o desempenho durante o jogo, usando as métricas esperadas apresentadas anteriormente (xG, xGA, $NSxG$ e $NSxGA$). Os aG serão deixados de fora porque foram concebidos com a finalidade de neutralizar o ambiente da partida, não sendo interessante ao objetivo da pesquisa.

3.1 Mínimos Quadráticos

Para estimar nosso modelo, usaremos o método de mínimos quadráticos ordinários (MQO). Como citam Gujarati & Porter (2011), a intenção é que, para um conjunto de n amostras de pares de Y e X, é necessário encontrar uma forma de que o somatório dos resíduos de uma regressão $\sum \hat{u}_i$ seja o menor possível.

Uma alternativa para essa situação seria o uso do critério dos mínimos quadráticos, tal que:

$$EQT = \sum \hat{u}_i^2 = f(\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_i), \tag{3.1}$$

onde o somatório total do quadrado dos resíduos (EQT) é uma função dos múltiplos parâmetros $\hat{\beta}_i$.

Em resumo, o método de MQO será capaz de encontrar os coeficientes que minimizam EQT, ajustando a reta da regressão para que esta apresente as menores distâncias quadráticas entre os valores observados e os calculados pelo modelo (Maia, 2017).

Assim, para uma regressão hipotética de parâmetros β_0 e β_1 e β_3, temos que:

$$\frac{\partial EQT}{\partial \beta_0} = 0 \quad \text{e} \quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}_1 - \hat{\beta}_2 \bar{X}_2, \tag{3.2}$$

$$\frac{\partial EQT}{\partial \beta_1} = 0 \quad \text{e} \quad \hat{\beta}_1 = \frac{\left(\sum y_i x_{1i}\right)\left(\sum x_{2i}^2\right) - \left(\sum y_i x_{2i}\right)\left(\sum x_{1i} x_{2i}\right)}{\left(\sum x_{1i}^2\right)\left(\sum x_{2i}^2\right) - \left(\sum x_{1i} x_{2i}\right)^2}, \tag{3.3}$$

$$\frac{\partial EQT}{\partial \beta_2} = 0 \quad \text{e} \quad \hat{\beta}_2 = \frac{\left(\sum y_i x_{2i}\right)\left(\sum x_{1i}^2\right) - \left(\sum y_i x_{1i}\right)\left(\sum x_{1i} x_{2i}\right)}{\left(\sum x_{1i}^2\right)\left(\sum x_{2i}^2\right) - \left(\sum x_{1i} x_{2i}\right)^2} \tag{3.4}$$

3.2 Modelo Linear

Utilizando os dados dispostos conforme tabela 2.1, o modelo construído para quantificarmos os efeitos do mando de campo sem a presença da torcida nos resultados dos jogos é dado por:

$$Saldo_{ijp} = \beta_{HFA} + \beta_1 Dif.SPI_{ijp} + \beta_2 Dif.xG_{ijp} + \beta_3 Dif.NSxG_{ijp} + \epsilon_{ijp}. \tag{3.5}$$

Em resumo, a equação expressa o $Saldo$, ou a diferença de gols, da partida p, identificada através do código ID em nossa base, em que os times i e j estão disputando.

Nesta ideia, será levada em consideração o referencial do time i, que será sempre o mandante, e o saldo será a quantidade de gols marcadas por este subtraída o *score*[1] do time j.

O modelo leva como *inputs* as diferenças de SPI, xG e $NSxG$ de cada um dos times na partida, onde o intercepto β_{HFA} será o responsável por captar o efeito da vantagem mando de campo (HFA) na variável resposta, por ser um efeito fixo e alheio à performance dos times. Como os SPI são dinâmicos, ou seja, mudam no decorrer do campeonato, é necessário que ambos os números inseridos no modelo sejam os de uma dada partida p. O erro será expresso por ϵ

A intenção é que o modelo tenha variáveis explanatórias tanto preditivas, como o SPI pré jogo, quando descritivas, através das medidas esperadas de xG e $NSxG$ que podem descrever o desempenho durante a partida.

Como visto, nosso modelo seria uma regressão linear múltipla, onde alguns dos pré requisitos para o bom funcionamento do modelo seriam que os dados devem ser distribuídos de maneira normal, com relação linear, livres de valores extremados e não tendo relações múltiplas entre variáveis independentes (Uyanik & Guler, 2013).

3.2.1 Performance do modelo linear

O modelo teve sua performance posta a prova, visando encontrar eventuais problemas no funcionamento dos estimadores e identificar se as hipóteses do Teorema de Gauss-Markov evidenciam que nossos estimadores são os Melhores Estimadores Lineares Não Viesados (BLUE), onde por não viesado entendesse por:

$$E(\hat{\beta})) = \beta, \tag{3.6}$$

e que:

$$Var(\hat{\beta})) = Var(\hat{\beta}')), \tag{3.7}$$

um estimador mais eficiente que $\hat{\beta}'$, qualquer que seja $\hat{\beta}'$ outro estimador de β (Maia, 2017).

Seguindo o modelo clássico de regressão linear, segundo (Maia, 2017), os pressupostos para que os parâmetros estimados sejam considerados BLUE seriam:

Tabela 3.1: Pressupostos - BLUE

Estimadores BLUE	
Índice	Descrição
1	Relação Linear
2	Valores de X são fixos em repetidas amostras
3	Esperança condicional dos erros igual a zero
4	Variabilidade dos erros constante
5	Erros não autocorrelacionados
6	Erros distribuídos normalmente

Fonte: O Autor (2022).

O item 6 da tabela 3.1 não é de fato um pressuposto, mas auxilia no funcionamento do modelo.

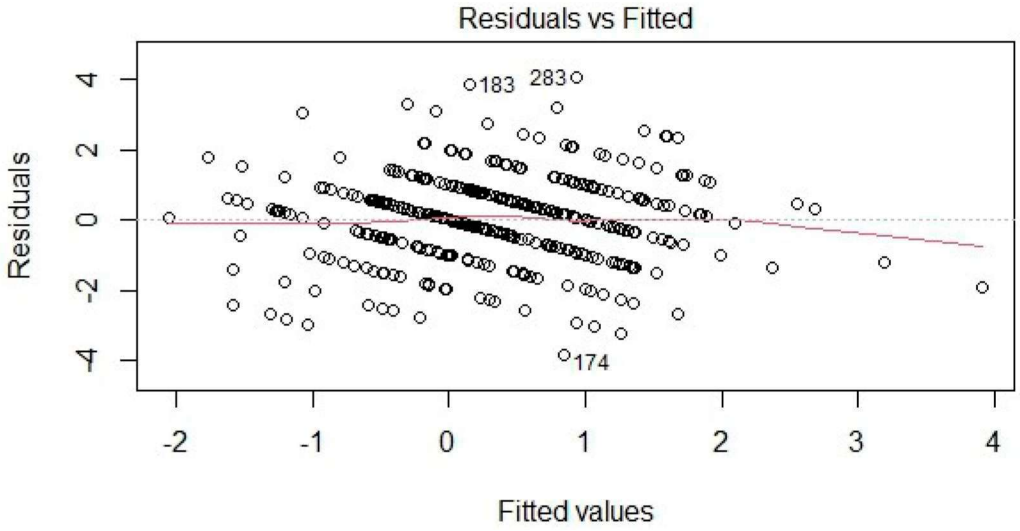
[1] Quantidade de gols.

3.2.1.1 Linearidade

Os gráficos apresentados nas seguintes sessões foram gerados com o intuito de identificar se nossa equação atende os pré requisitos citados nos capítulos anteriores.

O primeiro teste será o de linearidade. O gráfico a seguir (3.1) expressa a relação entre os valores respostas gerados e os resíduos do modelo. A hipótese de linearidade é respeitada quando identificado que a linha de referência seja o mais horizontalizada possível em torno do 0 (Lüdecke et al., 2021).

Figura 3.1: Teste de Linearidade



Fonte: O Autor (2022).

Como evidenciado, nosso modelo atendo a hipótese, já que os valores, em média, giram em torno do valor 0 no eixo Y, indicando uma relação linear. Sendo assim, a primeira hipótese do Teorema de Gauss-Markov é respeitada.

3.2.1.2 Variância e homocedasticidade

O próximo passo é a identificação do comportamento da variância. (Lüdecke et al., 2021) citam a importância de que os erros da regressão sejam identicamente distribuídos e com variância contínua, levando assim à um modelo homocedástico. Caso o contrário ocorra, o tal será definido como heterocedástico.

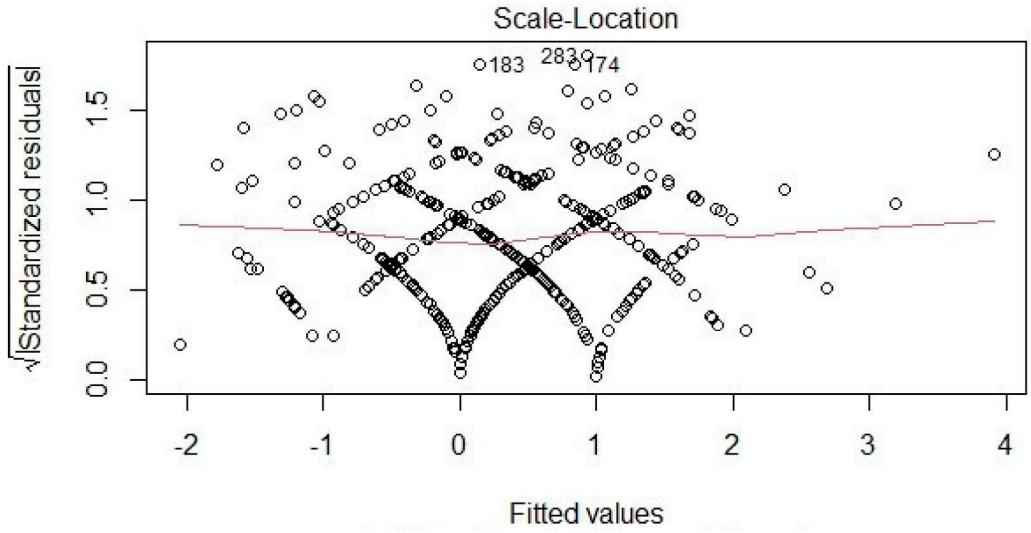
Para que os estimadores sejam BLUE, há a necessidade de variância constante de amostra, ou seja:

$$E(u_i^2) = \sigma^2, \tag{3.8}$$

Algumas causas para a heterocedasticidade seriam omissão de variáveis, a presença de *outliers*[2] ou uma má especificação na equação (Klein et al., 2015). Para verificar se há a presença deste problema, iremos realizar, primeiramente, uma análise gráfica e um teste estatístico.

[2]Observações fora do padrão.

Figura 3.2: Teste de Homocedasticidade



Fonte: O Autor (2022).

A figura 3.2 representa a relação entre as variáveis respostas e a raiz quadrática dos resíduos padrões do modelo construído no trabalho. Lüdecke et al. (2021) relatam que a presença de homogeneidade de variância é comprovada quando a linha de referencial fica horizontalizada. Logo, nossa regressão obtêm bom resultado na parte gráfica da análise.

O próximo passo será a execução de um teste capaz de rejeitar a hipótese de heterocedasticidade. Nossa escolha será pelo teste proposto por White (1980), uma dos mais utilizadas ferramentas de diagnóstico atualmente em trabalhos aplicados (Halunga et al., 2017).

Seja o Teste de White, para o modelo do presente trabalho, dado por:

$$\begin{aligned}\hat{e}i^2 = {} & \delta 0 + \delta_1 Dif.SPI_{ijp} + \delta_2 Dif.xG_{ijp} + \delta_3 Dif.NSxG + \delta_4 Dif.SPI_{ijp} Dif.xG_{ijp} \\ & + \delta_5 Dif.SPI_{ijp} Dif.NSxG_{ijp} + \delta_6 Dif.xG_{ijp} Dif.NSxG_{ijp} + \delta_7 Dif.SPI^2_{ijp} + \delta_8 Dif.xG^2_{ijp} \\ & + \delta_9 Dif.NSxG^2_{ijp} + u_{ijp},\end{aligned} \tag{3.9}$$

onde as hipóteses nula e alternativa são:

$$\begin{cases} H_0 : \delta_0 = \delta_1 = \delta_2 = \delta_3 = \delta_4 = \ldots = 0 & (3.10) \\ H_1 : \delta_k \neq 0, & (3.11) \end{cases}$$

e definido por:

$$LM = nR^2, \tag{3.12}$$

Para n amostras com um coeficiente R^2.

Foram alcançados os seguintes resultados, conforme tabela 3.2.

Sendo assim, para um nível de significância $\alpha = 0{,}05$, aceitamos a hipótese nula para a presença de homoscedasticidade.

Tabela 3.2: Teste de White

Teste de White	
Test Statistic	P-Value
0,57	0,7504

3.2.1.3 Teste de Multicolinearidade

Multicolinearidade é definida como uma situação em que múltiplas variáveis refletem uma variabilidade correlacionada, como menciona Voss (2005). Sendo assim, ausência de colinearidade significa que nenhum dos regressores pode ser expresso como uma combinação linear exata dos demais regressores do modelo.

Para mensurar o grau de multicolinearidade em nossa regressão, estaremos utilizando a métrica FIV, ou o Fator de Inflação da Variância. O FIV como um método para mostrar como a variância de um estimador é inflada pela presença da multicolinearidade. Ele é dado por:

$$FIV = \frac{1}{1 - R_{ab}^2}, \tag{3.13}$$

Onde a e b são duas variáveis explanatórias quaisquer de um modelo.

Aplicamos a FIV em nosso modelo para verificar a presença de multicolinearidade entre os *inputs* de nossa regressão, obtendo os seguintes resultados, conforme tabela 3.3.

Tabela 3.3: Teste de FIV

Teste de FIV		
$Dif.SPI$	$Dif.xG$	$Dif.NSxG$
1,482399	1,534667	1,668680

Fonte: O Autor (2022).

Belsley et al. (1980) sugerem que os níveis de FIV não devem ultrapassar o valor 10. Logo, os resultados do teste realizado revelam ausência, ou presença em níveis muito pequenos, de multicolinearidade.

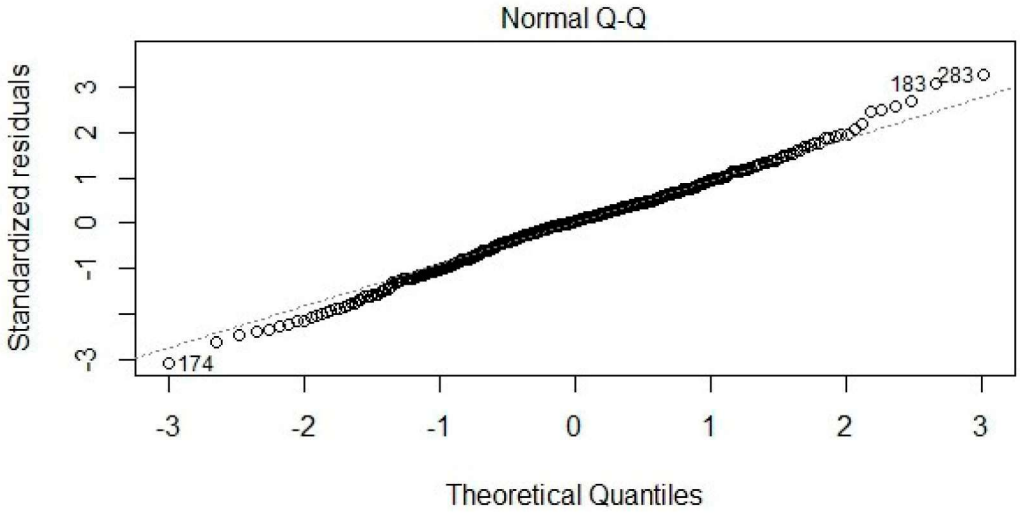
3.2.1.4 Normalidade dos resíduos

O próximo passo é checar se nosso modelo segue o pré requisito de normalidade dos resíduos. Assim como nos testes de homoscedasticidade, a análise será subdividida em uma parte gráfica e na performa de um teste estatístico comprobatório.

A figura 3.3 é um gráfico Q-Q, onde duas distribuições de probabilidade são comparados por meio de seus quantils (Wilk & Gnanadesikan, 1968). No caso, esta sendo comparados os quantils da distribuição da amostra com os de uma normal padronizada. Lüdecke et al. (2021) apontam que se os pontos estiverem concentrados sobre a linha diagonal, este seria um bom indício de normalidade residual, similar ao que é visto no gráfico anterior (3.3).

Sendo assim, para confirmar a impressão deixada pela análise gráfica, será realizado um teste de Shapiro-Wilk. Segundo Razali & Wah (2011), o teste de Shapiro-Wilk pode ser considerado o mais poderoso e preciso método de análise da normalidade dos resíduos obtidos por uma regressão, sendo dado por, como definido por Shapiro & Wilk (1965) pela equação 3.13.

Figura 3.3: Teste de Normalidade dos Resíduos



Fonte: O Autor (2022).

$$W = \frac{\left(\sum_{i=1}^{n} a_i y_i\right)^2}{\sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2}, \tag{3.14}$$

Onde y_i é a i^{th} estatística, $\bar{y}$ é a média amostral,

$$\boldsymbol{a}_i = (a_1, \cdots, a_n) = \frac{m^T V^{-1}}{\left(m^T V^{-1} V^{-1} m\right)^{1/2}}, \tag{3.15}$$

e $m = (m_1, \ldots, m_n)^T$ são os valores esperados das estatísticas de ordem de independente e variáveis randômicas distribuídas identicamente retiradas de uma normal padrão. V é a matriz de covariância destas estatísticas de ordem (Razali & Wah, 2011).

Os valores de W variam entre 0 e 1, onde pequenos levam à rejeição e altos à aceitação da hipótese de normalidade da amostra (Razali & Wah, 2011). As hipóteses serão:

$$\begin{cases} H_0 : \text{Dados Normalmente Distribuídos} & (3.16) \\ H_1 : \text{Dados Não Normalmente Distribuídos.} & (3.17) \end{cases}$$

Os resíduos da regressão do trabalho foram separados e postos à prova do teste descrito anteriormente, obtendo os seguintes resultados, conforme tabela 3.4.

Tabela 3.4: Teste de Shapiro-Wilk

Teste de Normalidade - Shapiro-Wilk	
W Statistic	P-Valor
0,99478	0,2307

Fonte: O Autor (2022).

Sendo assim, como indicado pelo alto valor de W, podemos aceitar a hipótese de que os resíduos do modelo elaborado são normalmente distribuídos.

3.2.1.5 Autocorrelação

O último item a ser checado para confirmar a qualidade dos parâmetros será o de autocorrelação dos resíduos. Como cita Maia (2017), por autocorrelação se subentende a associação entre os valores de uma mesma variável e é evidenciada facilmente quando os dados, neste caso os erros da regressão, podem ser dispostos de maneira espacial ou temporal, tal que:

$$\text{Cov}\left(e_t, e_{t+s}\right) = E\left(e_t e_{t-s}\right) \neq 0, \tag{3.18}$$

Maia (2017) ainda cita que algumas das fontes para a autocorrelação dos resíduos seriam a inércia dos ciclos, como no caso das séries temporais, as falhas de especificação do modelo e as defasagens.

A autocorrelação é particularmente importante pois apesar dos parâmetros continuarem consistentes e não viciados eles deixam de ser eficientes e significantes (Maia, 2017).

Para comprovar ou não a existência de autocorrelação nos resíduos do modelo do presente trabalho, será performado um teste de Durbin-Watson, um dos mais famosos e utilizados para a detecção serial.

Como formulado originalmente por Durbin & Watson (1950), o teste d é dado por:

$$d = \frac{\sum_{t=2}^{t=n}\left(\hat{u}_t - \hat{u}_{t-1}\right)^2}{\sum_{i=1}^{t=n}\hat{u}_t^2} \quad \text{e que} \quad d = 2(1-\rho), \tag{3.19}$$

Onde ρ é um coeficiente de autocorrelação serial, variando entre -1 e 1. A estatística d é um valor que varia entre 0 e 4, onde se $d = 2$, e por consequência $\rho = 0$, não há evidência de autocorrelação.

Neste teste, o coeficiente ρ será dado por:

$$\hat{\rho} = \frac{\sum \hat{u}_t \hat{u}_{t-1}}{\sum \hat{u}_t^2}, \tag{3.20}$$

e as hipóteses para o teste realizado serão:

$$\begin{cases} H_0 : \rho = 0 & (3.21) \\ H_1 : \rho \neq 0, & (3.22) \end{cases}$$

e também:

$$\begin{cases} H_0 : d = 2 & (3.23) \\ H_1 : d < 2. & (3.24) \end{cases}$$

Uma vez realizado o teste, foram obtidos os seguintes resultados, conforme tabela 3.5.

Sendo assim, para um nível de significância $\alpha = 0{,}05$, podemos aceitar a hipótese nula para a ausência de autocorrelação. Por consequência, podemos concluir que os parâmetros da regressão estimada cumprem todas as hipóteses do Teorema de Gauss-Markov, podendo assim serem classificados como BLUE.

Tabela 3.5: Teste de Durbin-Watson

Teste de Durbin-Watson	
d	P-Valor
2,1926	0,0612

Fonte: O Autor (2022).

3.3 Resultados Obtidos

Foi elaborada o tratamento dos dados e o modelo foi então posto a prova, obtendo os resultados numéricos expostos na tabela 3.6.

Tabela 3.6: Resultados do Modelo

	Variável Dependente:
	Diferença Entre Gols Marcados e Sofridos
$Dif.SPI$	0.022^{***}
	(0.006)
$Dif.xG$	0.791^{***}
	(0.072)
$Dif.NSxG$	-0.484^{***}
	(0.095)
HFA	0.231^{***}
	(0.068)
Observations	377
R^2	0.318
Adjusted R^2	0.312
Residual Std. Error	1.255 (df = 373)
F Statistic	57.872^{***} (df = 3; 373)
Nota:	*p<0.1; **p<0.05; ***p<0.01

Fonte: O Autor (2022).

Como evidenciado, em nosso modelo todas as variáveis foram consideradas significativas a um nível $\alpha = 0{,}05$. Da mesma forma, a estetística F apontou significância global dos parâmetros.

Com a intenção de aprofundar a análise, foram também elaboradas múltiplas combinações para o modelo usando as variáveis explanatórias da tabela 3.6, tentando entender se a omissão de algum *input* possa melhorar a performance da regressão.

Foram obtidos os seguintes resultados, expostos na tabela 3.7.

Para comparar os resultados das diferentes regressões possíveis, usaremos o Critério de Informação de Aikaike, ou AIC (Akaike, 1973), dado pela equação 3.24.

$$AIC = 2k - 2\ln(\hat{L}), \tag{3.25}$$

Tabela 3.7: Resultados das Combinações de Regressões

	Variável Dependente:			
	Diferença Entre Gols Marcados e Sofridos.			
	(1)	(2)	(3)	(4)
$Dif.SPI$	0.022^{***}	0.011^{*}	0.039^{***}	
	(0.006)	(0.006)	(0.007)	
$Dif.xG$	0.791^{***}	0.641^{***}		0.857^{***}
	(0.072)	(0.068)		(0.071)
$Dif.NSxG$	-0.484^{***}		-0.061	-0.354^{***}
	(0.095)		(0.099)	(0.090)
HFA	0.231^{***}	0.156^{**}	0.336^{***}	0.186^{***}
	(0.068)	(0.069)	(0.078)	(0.068)
Observations	377	377	377	377
R^2	0.318	0.270	0.099	0.293
Adjusted R^2	0.312	0.266	0.094	0.289
Residual Std. Error	1.255 (df = 373)	1.296 (df = 374)	1.440 (df = 374)	1.276 (df = 374)
F Statistic	57.872^{***} (df = 3; 373)	69.206^{***} (df = 2; 374)	20.569^{***} (df = 2; 374)	77.426^{***} (df = 2; 374)
AIC	1247.090	1270.461	1349.827	1258.556
Nota:				*p<0.1; **p<0.05; ***p<0.01

Sendo k o número de variáveis e $\hat{L}$ o valor máximo da função de verossimilhança. Quando comparado dois ou mais modelos, é eligido como melhor o que tem um resultado de AIC menor. O critério é responsável por recompensar modelos estatísticos pela sua qualidade levando em consideração a mesma base de dados (Akaike, 1973).

Como evidenciado pela tabela 3.7, a combinação original de variáveis obteve o menor resultado na métrica, e portando o melhor. Sendo assim, esta será eligida para a análise dos resultados.

4 Análise dos Resultados

O modelo obteve um β_{HFA} de 0,231, o que nos levaria a concluir que a vantagem do mando de campo continuou nos jogos do Campeonato Brasileiro mesmo sem a presença de torcida. Imaginando um cenário hipotético onde dois times exatamente iguais, jogassem dez partidas com desempenhos idênticos, o time mandante venceria pelo menos duas delas justamente pelo fato de ser o dono do estádio, mesmo que não haja torcida.

O resultado é muito surpreendente, pois é de se pensar, por lógica, que a ausência de torcida reduziria o efeito mando praticamente a zero, o que não é visto. O modelo também ressaltou que o desempenho dentro do jogo é capaz de determinar melhor o resultado da partida do que o momento dos times *ex-ante*, já que:

$$|\beta_{Dif.xG_ijp}| > |\beta_{Dif.NSxG_ijp}| > |\beta_{Dif.SPI_ip}|, \tag{4.1}$$

Mostrando que a magnitude nas diferenças de xG e $NSxG$ são maiores do que a de SPI.

O modelo apontou um R^2 de 0,318 e um $\bar{R}^2$ de 0,312, tendo assim boa capacidade de determinação. A tabela ANOVA da regressão elaborada é dada em 4.1

Tabela 4.1: Tabela Anova

Statistic	N	Mean	St. Dev.	Min	Max
Df	4	94.000	186.000	1	373
Sum Sq	4	215.218	252.016	40.894	587.440
Mean Sq	4	68.751	62.770	1.575	148.017
F value	3	57.872	34.204	25.966	93.985
Pr(>F)	3	0.00000	0.00000	0.000	0.00000

Fonte: O Autor (2022)

O presente trabalho conseguiu, desta forma, evidenciar que a vantagem do mando de campo continuou existindo no Brasileirão 2020 mesmo com a pandemia do COVID-19, que impôs medidas restritivas a presença de torcedores nos estádios brasileiros.

O próximo passou seria avaliar se a HFA se manteve nos níveis similares aos de campeonatos anteriores (com torcida) ou se acabou decrescendo. Sendo assim, utilizando o mesmo modelo, regredimos os dados das temporadas 2018 e 2019, que tiveram uma amostragem de 380 e 355 jogos, respectivamente.

Foram obtidos os seguintes resultados (4.2)

Tabela 4.2: Resultados - Mando de Campo

Resultados - HFA		
$\beta_{HFA_{2018}}$	$\beta_{HFA_{2019}}$	$\beta_{HFA_{2020}}$
0,445	0,338	0,231

Fonte: O Autor (2022)

Em resumo, a presente monografia concluiu que ao mesmo tempo que a vantagem do mandante ainda está presente, ela diminuiu consideravelmente em relação aos níveis de

2018 e 2019. Os resultados apontam que ainda que a ausência de torcedores não coloque os times participantes em um patamar de igualdade, ela é responsável por reduzir o valor do mando de campo à um nível muito mais próximo disso.

5 Conclusão

Com a pandemia do COVID-19, a comunidade de análises estatísticas esportivas levantou diversas hipóteses de como a ausência de torcida afetaria as competições e campeonatos, seja de qual for a modalidade, e, mais especificamente de casos como o do futebol, se ainda faria diferença ser o mandante de uma partida.

O presente trabalho evidenciou que apesar da ausência de torcedores nas arquibancadas em jogos do Campeonato Brasileiro durante a pandemia, é possível concluir que a equipe mandante ainda tem certa vantagem em relação ao visitante. Se comprova também a importância do momentos dos times pré partida (através do SPI) e o desempenho durante esta (xG e $NSxG$) no resultado final.

É incerto qual seria a fonte desta vantagem, porém os dados apontam significância nos resultados obtidos. É evidenciado também uma acentuada redução nos níveis de HFA em relação à jogos pré pandemia, o que nos leva a concluir que mesmo a vantagem do mandante ainda estando presente, ela é muito menos influente quando comparada em jogos com torcida.

REFERÊNCIAS

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *2nd International Symposium on Information Theory.*, 267–281.

Barros Filho, C. R. (2002). Economia brasileira contemporânea.

Belsley, D., Kuh, E., & Welsch, R. (1980). Regression diagnostics: Identifying influential data and sources of collinearity. *John Wiley and Sons.*

Benz, L. & Lopez, M. (2021). *Estimating the change in soccer's home advantage during the Covid-19 pandemic using bivariate Poisson regression.* PhD thesis.

Boice, J. (2020). How our club soccer predictions work. https://fivethirtyeight.com/methodology/how-our-club-soccer-predictions-work/. Accessado em 04/08/2022.

Bryson, A., Dolton, P., Reade, J., Schreyer, D., & Singleton, C. (2020). *Causal effects of an absent crowd on performances and refereeing decisions during covid-19.* PhD thesis.

Cueva, C. (2020). *Animal spirits in the beautiful game. testing social pressure in professional football during the covid-19 lockdown.* PhD thesis.

Dilger, A. & Vischer, L. (2020). No home bias in ghost games. *Institute for Organisational Economics.*

Durbin, J. & Watson, G. S. (1950). Testing for serial correlation in least squares regression, i. *Biometrika.*, *37*(3), 409–428.

Eggels, H. (2016). *Expected goals in soccer explaining match results using predictive analytics.* PhD thesis.

Endrich, M. & Gesche, T. (2020). Home-bias in referee decisions: Evidence from "ghost matches" during the covid19-pandemic. *Economics Letters*, *197*.

Ferraresi, M. & Gucciardi, G. (2020). Team performance and audience: experimental evidence from the football sector. Technical report, Siep.

Fischer, K. & Haucap, J. (2020). Does crowd support drive the home advantage in professional soccer? evidence from german ghost games during the covid-19 pandemic. *Journal of Sports Economics*, *22*(8), 982–1008.

Furtado, C. (2020). *Formação econômica do Brasil.* Companhia das Letras.

Garicano, L., Palacios-Huerta, I., & Prendergast, C. (2005). Favoritism under social pressure. *Review of Economics and Statistics*, *87*(2), 208–216.

Gujarati, D. N. & Porter, D. C. (2011). *Econometria básica-5.* Amgh Editora.

Halunga, A., Orme, C., & Yamagata, T. (2017). A heteroskedasticity robust breusch-pagan test for contemporaneous correlation in dynamic panel data models. *Journal of Econometrics*, 209–230.

Klein, A., Gerhard, C., Buchner, D., Diestel, S., & Schermelleh-Engel, K. (2015). The detection of heteroscedasticity in regression models for psychological data. *Psychological Test and Assessment Modeling*, *58*, 542–XXX.

Krawczyk, M. & Strawinski, P. (2020). Home advantage revisited. did covid level the playing fields? Technical report, Sciendo.

Leitner, M. & Richlan, F. (2020). No fans-no home advantage: Sport psychological effects of missing supporters on football teams in european top leagues.

Lewis, M. (2003). *Moneyball: The Art of Winning an Unfair Game*. W. W. Norton Company.

Lüdecke, D., Ben-Shachar, M., Patil, I., & Makowski, D. (2021). Extracting, computing and exploring the parameters of statistical models using r. *JOSS*.

Lüdecke, D., Ben-Shachar, M. S., Patil, I., Waggoner, P., & Makowski, D. (2021). performance: An R package for assessment, comparison and testing of statistical models. *Journal of Open Source Software*, *6*(60), 3139.

Maia, A. (2017). *Econometria: conceitos e aplicações*. Editora Saint Paul.

McCarrick, D., Bilalic, M., Neave, N., & Wolfson, S. (2021). Home advantage during the covid-19 pandemic in european football. *Elsevier*, *56*.

Morgulev, E., Azar, O., & Lidor, R. (2018). Sports analytics and the big-data era. *Int J Data Sci Anal*, *5*(1), 213–222.

Moskowitz, T. & Wertheim, L. (2012). Scorecasting: The hidden influences behind how sports are played and games are won. *Three Rivers Press*.

Naicker, V. (2021). How math and data science made liverpool the best team on the planet. https://medium.com/the-spekboom/how-math-and-data-science-made-liverpool-the-best-team-on-the-planet-a72d50b325. Accessado em 04/08/2022.

Pettersson-Lidbom, P. & Priks, M. (2010). Behavior under social pressure: Empty italian stadiums and referee bias. *Economics Letters*, *108*(2), 212–214.

Rathke, A. (2017). An examination of expected goals and shot efficiency in soccer. *Journal of Human Sport and Exercise*, *12*(2), 514–529.

Razali, A. & Wah, Y. (2011). Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. *Journal of Statistical Modeling and Analytics.*, *2*(1), 21–33.

Reade, J., Schreyer, D., & Singleton, C. (2020). Echoes: what happens when football is played behind closed doors?. *Economic Inquiry*.

Robberechts, P. & Davis, J. (2020). How data availability affects the ability to learn good xg models. *Communications in Computer and Information Science*, *1324*.

Sanchez, A. & Lavin, J. (2020). Home advantage in european soccer without crowd. *Soccer and Society*, 1–14.

Schoenfeld, B. (2019). How data (and some breathtaking soccer) brought liverpool to the cusp of glory. https://www.nytimes.com/2019/05/22/magazine/soccer-data-liverpool.html. Accessado em 04/08/2022.

Scoppa, V. (2020). Social pressure in the stadiums: Do agents change behavior without crowd support? *Journal of Economic Psychology*, *82*.

Shapiro, S. & Wilk, M. (1965). An analysis of variance test for normality (complete samples). *Biometrika.*, *52*(4), 591–611.

Sors, F., Grassi, M., Agostini, T., & Murgia, M. (2020). The sound of silence in association football: Home advantage and referee bias decrease in matches played without spectators. *European journal of sport science*, 1–21.

Unkelbach, C. & D., M. (2010). Crowd noise as a cue in referee decisions contributes to the home advantage. *Journal of Sport and Exercise Psychology*, *32*(4), 483–498.

Uyanik, G. & Guler, N. (2013). A study on multiple linear regression analysis. *Elsevier*, *106*, 234–240.

Voss, S. (2005). Multicollinearity. *Encyclopedia of Social Measurement*, 759–770.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *JSTOR*, *48*(4).

Wilk, M. & Gnanadesikan, R. (1968). Probability plotting methods for the analysis of data. *Biometrika Trust.*, *55*(1), 1–17.