Caracterização de Estratégias de Futebol com base na Frequência, Importância e Efetividade de Jogadas

Characterizing Soccer Strategies based on Moves' Frequency, Importance and Effectiveness*

Gabriel Valadao[1], João L. L. Gonçalves[1], João L. L. Megale[1], Vinicius M. Paula[1], Hugo Rios-Neto[1], Adriano C. M. Pereira[1], Wagner Meira Jr.[1]

[1]Sports Analytics Lab – Speed
Departamento de Ciência da Computação
Universidade Federal de Minas Gerais

{gabriel.valadao, joao.lucas, joaomegale}@dcc.ufmg.br, afssvini@ufmg.br

{hugoriosneto, adrianoc, meira}@dcc.ufmg.br

Abstract. *Understanding and predicting soccer results is a challenge due to the complexity of the game itself, the number and diversity of players involved and even external factors that are not always qualifiable or quantifiable. On the other hand, there is a significant effort to train teams and prepare them to act and react appropriately to a variety of scenarios, which indicates the existence of strategies. This paper aims to characterize these football strategies, focusing on how the team behaves collectively. We define football strategy as a set of moves that must be frequent, important and effective, which are normally conflicting criteria. We propose a methodology to identify these strategies, which is implemented and evaluated using real data from league seasons. The results show that our methodology allows characterizing the strategies of different teams.*

Resumo. *Entender e prever resultados em futebol é um desafio pela complexidade do jogo em si, pelo número e diversidade dos jogadores envolvidos e mesmo por fatores externos que nem sempre são qualificáveis ou quantificáveis. Por outro lado, há todo um esforço no sentido de treinar os times e prepará-los para agir e reagir adequadamente a uma variedade de cenários, o que indica a existência de estratégias. Este artigo tem por objetivo caracterizar essas estratégias de futebol, com foco em como o time se comporta coletivamente. Estratégia de futebol é definida como um conjunto de jogadas que devem ser frequentes, importantes e efetivas, que são critérios normalmente conflitantes. É proposta uma metodologia para identificar essas estratégias, a qual é implementada e avaliada usando dados reais de temporadas de ligas. Os resultados mostram que a metodologia proposta permite caracterizar estratégias dos diferentes times.*

*Os quatro primeiros autores contribuiram igualmente para o trabalho.

1. Introdução

A análise e a modelagem de dados no futebol têm ganhado atenção significativa recentemente, nas esferas públicas e acadêmicas, à medida que equipes e analistas buscam obter um entendimento mais preciso e detalhado do jogo [Harper 2021][Fernández et al. 2021]. Uma das principais tarefas nesta área é a detecção e análise do estilo de jogo, bem como os pontos fortes e fracos dos adversários. Ao identificar essas características, as equipes podem desenvolver estratégias eficazes e tomar decisões informadas, buscando uma vantagem competitiva.

No domínio da análise de dados no futebol, os dados de eventos desempenham um papel crucial, fornecendo informações detalhadas sobre os eventos que ocorrem durante as partidas, como passes, dribles, chutes, faltas, cartões, substituições, início e fim de tempo, entre outros [Pappalardo et al. 2019]. No escopo deste trabalho, define-se ação por eventos com bola deliberados que jogadores realizam durante a partida, como passes, dribles, cruzamentos e chutes [Decroos et al. 2019]. Ademais, define-se jogada por uma sequência de ações ofensivas consecutivas, realizadas por uma mesma equipe, do momento em que obtiveram a bola ao momento que a perderam.

Estudos anteriores, como o realizado por [McCarthy et al. 2023], abordaram a detecção do estilo de jogo, pontos fortes e fracos dos adversários, analisando as sequências de passes frequentes que levam a oportunidades de gol. No entanto, essa abordagem tem limitações que este trabalho busca superar. Primeiramente, a abordagem anterior concentrou-se principalmente nos passes e negligenciou outros tipos de ações que as equipes usam para criar oportunidades de gol. Ao excluir essas ações da análise, informações importantes sobre as capacidades de ataque de uma equipe podem passar despercebidas.

Em segundo lugar, o método anterior filtrou as subsequências que não terminaram com um chute com uma Expectativa de Gol (xG) [Ensum et al. 2004][Green 2012] de pelo menos 0,33. A métrica xG é uma estatística de futebol que estima a probabilidade de uma finalização resultar em gol com base em vários fatores, como sua localização, se houve assistência, o tipo de finalização, entre outros. Embora o xG capture sequências que têm probabilidade de resultar em um gol, ela desconsidera sequências que poderiam ter tido um alto potencial de marcar, mas foram interrompidas pela perda da posse de bola perto do momento de finalização. Ao excluir essas sequências, a análise pode deixar de identificar jogadas típicas em que a ameaça de ataque de uma equipe por pouco não resultou em uma finalização perigosa.

Para tentar solucionar essas limitações, este trabalho propõe uma nova abordagem para a construção de jogadas e a filtragem de jogadas consideradas como oportunidades de gol. Em primeiro lugar, a proposta inclui a construção de sequências de ações com base nos agrupamentos obtidos a partir do arcabouço SoccerMix[Decroos et al. 2020], que utiliza modelos de mistura para agrupar diferentes tipos de ações com base em sua localização e direção. Isso enriquece a representação das sequências ao incluir não apenas a localização, mas também o tipo de ação e direção, quando comparado ao método em [McCarthy et al. 2023].

Em segundo lugar, a proposta utiliza as estimativas de probabilidade derivadas do arcabouço VAEP (Valuing Actions by Estimating Probabilities) [Decroos et al. 2019] como limiar para identificar chances de gol. Essa abordagem considera a probabilidade

de uma ação resultar em um gol nas próximas 10 ações, ao invés de focar apenas no xG dos chutes, mantendo-se agnóstica em relação ao resultado específico de uma sequência, mesmo que ela não termine em um chute.

Este artigo apresenta uma abordagem mais generalista em relação a [McCarthy et al. 2023], que busca minerar sequências de ações frequentes para a caracterização do estilo de jogo, pontos fortes e pontos fracos de times de futebol. Ao explorar os agrupamentos do SoccerMix e as estimativas de probabilidade do VAEP, a análise resultante captura uma gama mais ampla de ações e suas co-ocorrências.

O restante desse artigo está organizado da seguinte forma: a Seção 2 apresenta uma revisão dos trabalhos relacionados na análise de futebol. A Seção 3 contextualiza as motivações do trabalho sob o contexto prático do futebol. A Seção 4 detalha a metodologia e descreve os dados utilizados em nossa análise. A Seção 5 apresenta os resultados de nossos experimentos e discute suas implicações e desdobramentos. Por fim, a Seção 6 conclui o artigo e destaca caminhos para pesquisas futuras sobre o tema.

2. Trabalhos Relacionados

Nesta seção, apresenta-se uma visão geral de trabalhos prévios relevantes que se concentram na análise de sequências de ações no futebol. São apresentadas e discutidas três referências principais, destacando suas metodologias, similaridades e diferenças em relação à nossa abordagem.

[Decroos et al. 2018] explora o problema da detecção automática de táticas a partir de dados de eventos, realizando a análise de sequências de ações executadas pelas equipes quando estão com a posse de bola. Sua abordagem envolve o agrupamento de sequências contíguas de ações, a ordenação dos agrupamentos com base em sua relevância esperada para o usuário, a busca por padrões frequentes de cada agrupamento e desenvolvimento de uma função de priorização para ordenar os padrões descobertos. Uma distinção notável em relação à nossa abordagem é que eles agrupam sequências inteiras antes de realizar a mineração de padrões frequentes. Em contraste, optamos por agrupar ações individuais e construir sequências a partir dos grupos atribuídos às ações. Uma limitação identificada em seu trabalho é que o agrupamento das sequências, prévio à busca por padrões frequentes, pode fazer com que alguns padrões que possam vir a ser descobertos acabem sendo atribuídos a diferentes grupos e não sejam identificados.

Em [Malqui et al. 2019], os autores abordam a análise visual da complexidade das sequências de passes. Especificamente, eles se concentram em sequências de três passes, examinando as combinações de três ou quatro jogadores que participam dessa troca de passes e agrupando as trajetórias dos passes nessas sequências. Ao contrário de nossa abordagem, eles consideram apenas sequências de passes com comprimento três, restringindo a análise a um tamanho específico de sequência. Essa restrição dificulta a exploração de padrões de sequências mais longas e ricas que possam existir.

[McCarthy et al. 2023] concentra-se em identificar subsequências frequentes de passes que levam a oportunidades de marcar gols. Os autores constroem sequências de passes onde cada item da sequência corresponde a um passe e seu identificador de item é uma zona do campo. Eles filtram as subsequências que não resultam em um chute com um valor esperado de gols (xG) [Ensum et al. 2004] de pelo menos 0,33, que é considerado uma medida razoável de sequências com potencial de marcar gols. Em seguida,

é realizada a mineração de padrões sequenciais para revelar em quais zonas as equipes executam passes que levam a chances de perigo. Esse trabalho está intimamente alinhado com nosso objetivo de pesquisa. Consideramos sua abordagem a mais aplicável para identificar os padrões de jogadas de perigo de equipes. Isso se deve, em parte, por não agruparem sequências antes de realizar a mineração de padrões e não restringirem o tamanho das sequências.

3. Contexto

O objetivo deste trabalho é qualificar e quantificar jogadas típicas dos times e sua efetividade, como uma forma de sintetizar as estratégias e auxiliar no processo de aperfeiçoamento tático, tanto de ataque quanto de defesa. Essas jogadas típicas são sequências, não necessariamente contíguas de ações ou abstrações de ações. Nesse trabalho, são avaliadas jogadas típicas em três dimensões, a saber:

Frequência: A frequência de uma jogada típica é importante no sentido de quantificar a sua significância na estratégia do time. Jogadas frequentes permitem identificar pontos fortes de um time ou pelo menos ações recorrentes. Identificar jogadas frequentes implica em conseguir visualizar e recuperar jogadas que, mesmo que não importantes ou eficientes, acontecem a todo momento. Conseguir tirar proveito estratégico dessas jogadas típicas do adversário pode significar oportunidades para aplicar contra-estratégias. Como discutido na Seção 4.5, o suporte foi adotado como medida de frequência das jogadas.

Importância: A importância de uma jogada típica se refere ao quanto ela não é acidental, ou seja, aconteceu simplesmente como consequência do quão frequentemente cada jogador realiza uma ação com a bola ou a bola estar em uma dada região do campo. A identificação de jogadas importantes busca distinguir aqueles períodos do jogo quando nada realmente importante acontece de momentos onde jogadas mais elaboradas são realizadas. Esta métrica possibilita identificar comportamentos do time que podem surpreender o adversário. Como discutido na Seção 4.5, o interesse (*lift*) das jogadas foi adotado como métrica de importância, assim como o comprimento das jogadas (Seção 4.6).

Efetividade: A efetividade de uma jogada típica indica o quanto ela levou a uma vantagem competitiva, como por exemplo um gol. Nesse caso, realmente queremos avaliar o quanto a jogada típica tende a impactar o resultado do jogo ou não. As jogadas efetivas possibilitam identificar e avaliar que tipo de jogadas de um time - em termos de localização, sequência, direção, ações - criam mais chance de gol. A Seção 4.3.3 apresenta o PScore como medida de efetividade.

As diferentes métricas avaliadas podem ser combinadas, mas devido à variedade de pontuação por ações e alta complexidade do futebol, dificilmente uma jogada típica pontuará bem nas três métricas. Assim, existe um compromisso entre elas, de modo que as jogadas com valores mais altos em uma das métricas podem não pontuar tão bem nas outras. No entanto, encontrar sequências que, de maneira combinada, pontuam melhor em duas ou mais métricas é algo relevante no contexto de caracterização de estratégias de futebol. Assim, investigar formas de analisar conjuntamente tais dimensões ou critérios é também um dos objetivos deste trabalho.

4. Metodologia

Nesta seção descreve-se a metodologia para identificação das jogadas típicas, a qual é descrita em detalhes nas seções a seguir. Essa metodologia é sintetizada na Figura 1 e será descrita em detalhes nas subseções a seguir.

4.1. Dados de Eventos

Neste trabalho, foram utilizados os dados de eventos disponibilizados em [Pappalardo et al. 2019], que são a maior coleção aberta de registros de futebol já lançada, contendo todos os eventos espaciais-temporais (passes, chutes, faltas, etc.) que ocorreram durante cada partida por uma temporada inteira de várias competições de futebol proeminentes. As competições incluem: o campeonato inglês, o espanhol, o italiano, o francês, o alemão e a Liga dos Campeões. Cada evento da partida contém informações sobre sua posição, tempo, resultado, jogador e características.

Neste trabalho, foram utilizados somente dados do campeonato inglês por motivos de simplicidade e cobertura suficiente para demonstrar diversas aplicações e usos. Definitivamente, a metodologia poderia ser replicada e escalada para toda o conjunto de dados, no caso de buscar caracterizar times de outras ligas, ou em jogos europeus.

4.2. SPADL

Primeiramente, os dados de eventos são convertidos para a representação SPADL (Soccer Player Action Description Language) [Decroos et al. 2019], que padroniza as taxonomias e atributos armazenados de diferentes provedores de dados de eventos no futebol para usar em análises derivadas e como entrada de algoritmos de aprendizado de máquina. Para isso, foi utilizado o pacote SoccerAction [1]. Apesar de ser possível trabalhar com os dados de eventos originais para a construção das jogadas, o SPADL é suficiente para essa tarefa e, de qualquer forma, será necessário para os enriquecimentos das jogadas que realizamos. Dessa forma, decidimos realizar a conversão ao SPADL anteriormente à construção das jogadas. O SPADL armazena informações como identificadores do evento, jogo, time e jogador da ação, tempo de início da ação, tipo da ação, coordenadas x e y e parte do corpo com qual foi realizada a ação (pé ou cabeça). Há duas variações do SPADL, a original e uma variação chamada Atomic-SPADL, que omite o atributo do resultado da ação. O Atomic-VAEP foi escolhido por ter sido usado na implementação do SoccerMix [2] e por ter sido demonstrado ser mais confiável que o SPADL original para o arcabouço do VAEP [Davis et al. 2022].

4.3. Enriquecimento das Ações

Como mencionado anteriormente, tendo como referência o SPADL, cada ação pode ser associada a diferentes informações, refletindo propriedades da ação e do contexto da jogada. Em particular, são considerados dois tipos de enriquecimento: localização e potencial ofensivo. Localização tem por objetivo agrupar ações que sejam próximas e/ou semelhantes, habilitando a detecção de jogadas típicas. Em termos de localização, são considerados a utilização de uma grade pré-definida ou agrupamentos SoccerMix. O potencial ofensivo busca quantificar a chance de gol e este trabalho utiliza o VAEP. A seguir são descritas cada uma dessas estratégias de enriquecimento.

[1] https://github.com/ML-KULeuven/socceraction

[2] https://github.com/ML-KULeuven/soccermix/blob/master/notebooks/1-load-and-convert-statsbomb-data.ipynb

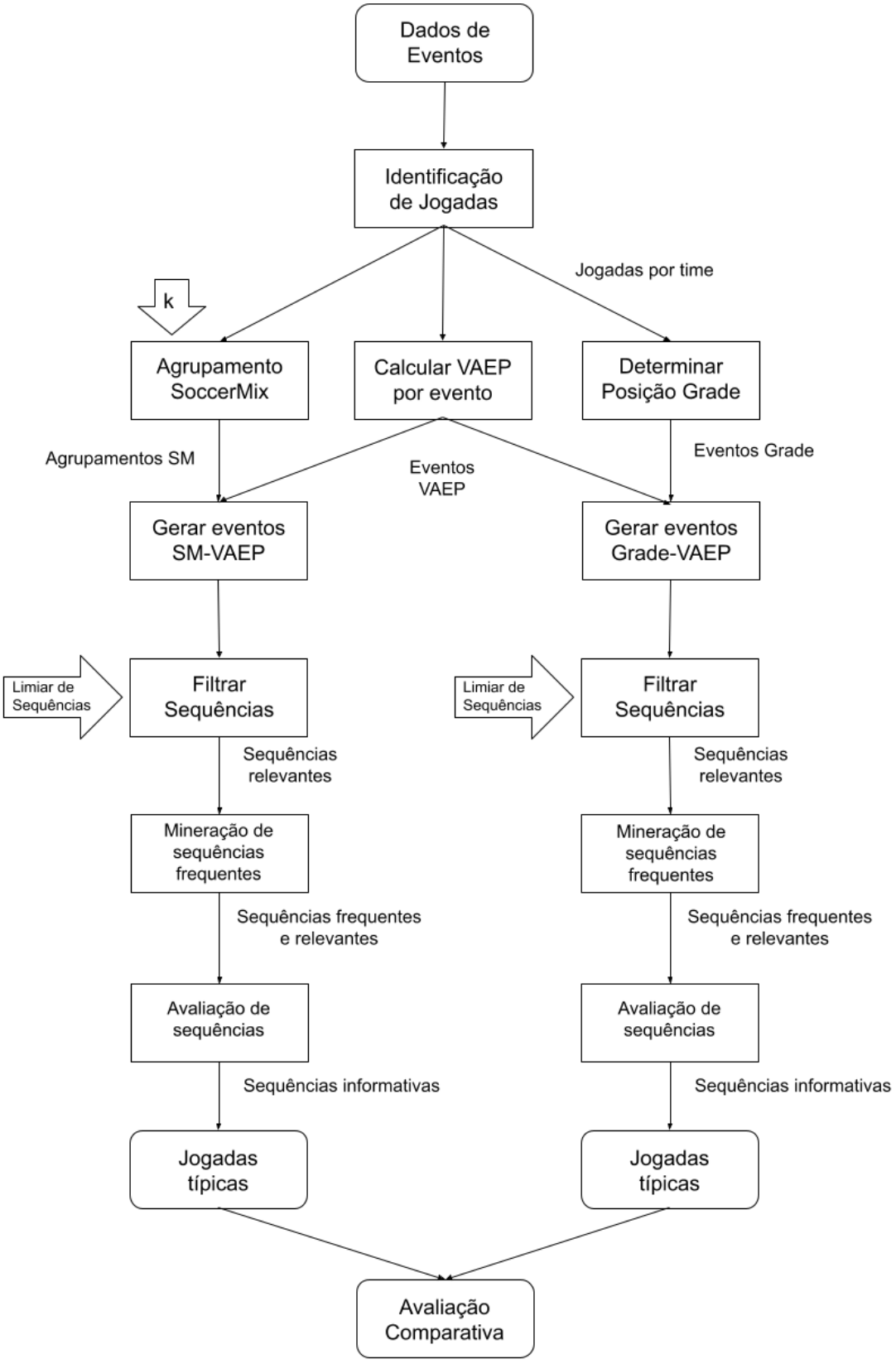

Figura 1. Metodologia

4.3.1. Localização com base na Posição na Grade

Uma estratégia simples para agrupar as localizações de ações que jogadores realizam é dividir o campo em zonas pré-definidas, à semelhança de uma grade, e considerar a localização de cada evento como sendo o centro da zona onde ele ocorreu [McCarthy et al. 2023]. Desta forma, eventos que ocorreram próximos serão considerados como tendo a mesma localização, facilitando a consolidação dessa informação.

Nesse trabalho é adotada a mesma grade do trabalho [McCarthy et al. 2023], que consiste de 28 células, numeradas sequencialmente entre 1 e 28. Utilizando as coordenadas x e y das ações presentes na representação Atomic-SPADL, é identificada a célula onde a ação ocorreu.

4.3.2. Localização usando Agrupamentos SoccerMix

O SoccerMix [Decroos et al. 2020] propõe agrupar, através de modelos de misturas, ações que se assemelhem em relação à localização e direção, por tipo de ação. Por utilizar de agrupamento suave, o método possibilita a representação probabilística de ações do jogo. Ou seja, cada ação pode ser associada à distribuição de probabilidade de que a ação pertença a cada um dos agrupamentos aprendidos pelo algoritmo. Originalmente, o SoccerMix foi usado, principalmente, para o problema de capturar o estilo de jogo de um jogador, ao agregar as probabilidade que as ações pertençam a cada um dos agrupamentos.

De maneira sucinta, o SoccerMix agrupa hierarquicamente, com modelos de misturas, em duas etapas. Na primeira etapa, para cada tipo de ação, um modelo de misturas de Gaussianas, que recebe como entrada coordenadas (x, y), é treinado. O resultado da primeira etapa são localizações prototípicas de onde ações podem ocorrer. Na segunda etapa, para cada localização prototípica descoberta na primeira etapa, é ajustado um modelo de mistura Von Mises, a fim de aprender as direções prototípicas de ações que são realizadas em cada uma das localizações. Por fim, a quantidade de componentes para cada localização e direção prototípicas são escolhidas através da formulação de um problema de otimização linear inteiro que visa maximizar o *Bayesian Information Criterion* (BIC) total [3] do conjunto de modelos de mistura escolhidos.

Nesse trabalho foram utilizadas apenas as localizações prototípicas, sem as direções, por elas criarem uma distinção por tipo de ação suficientemente granular para o objetivo do trabalho, que é a construção de sequências que levem em conta não somente a localização de ações, como seu tipo. Não foram incluídos os agrupamentos de direção para não aumentar significativamente a quantidade de possíveis atribuições a agrupamentos.

4.3.3. Valoração de ações usando VAEP

VAEP (*Valuing Actions by Estimating Probabilities*) [Decroos et al. 2019] é um arcabouço para valorar ações de jogadores de futebol. Valorar uma ação para uma equipe então requer avaliar a mudança na probabilidade tanto de marcar como de sofrer um gol

[3] https://github.com/ML-KULeuven/soccermix/blob/master/notebooks/2-create-mixture-models.ipynb

como resultado da ação a_i, que move o jogo do estado S_{i-1} para o estado S_i. Mais especificamente, a mudança na probabilidade da equipe x marcar um gol pode ser calculada como:

$$\Delta P_{\text{scores}}(a_i, x) = P_{\text{scores}}(S_i, x) - P_{\text{scores}}(S_{i-1}, x) \tag{1}$$

onde $P_{\text{scores}}(S_i, x)$ é a probabilidade do time x marcar um gol, à partir do estado de jogo S_i, dentro das próximas 10 ações. $\Delta P_{\text{scores}}(a_i, x)$ será positivo quando uma ação aumentar a probabilidade de resultar em gol para o time x. Similarmente podemos definir a mudança na probabilidade de um time x conceder um gol como:

$$\Delta P_{\text{concedes}}(a_i, x) = P_{\text{concedes}}(S_i, x) - P_{\text{concedes}}(S_{i-1}, x) \tag{2}$$

onde $P_{\text{concedes}}(S_i, x)$ é a probabilidade do time x sofrer um gol, à partir do estado de jogo S_i, dentro das próximas 10 ações. $\Delta P_{\text{concedes}}(a_i, x)$ será positivo quando uma ação aumentar a probabilidade de resultar em gol contra o time x. Seja y o adversário de x, note que $P_{\text{scores}}(S_i, x) = P_{\text{concedes}}(S_i, y)$ e $P_{\text{scores}}(S_i, y) = P_{\text{concedes}}(S_i, x)$.

Combinando as duas equações acima, temos na Equação 3 a definição do VAEP:

$$V(a_i, x) = \Delta P_{\text{scores}}(a_i, x) + (-\Delta P_{\text{concedes}}(a_i, x)) \tag{3}$$

onde o sinal negativo antes de $\Delta P_{\text{concedes}}(a_i, x)$ faz com que uma redução na probabilidade de sofrer um gol seja avaliada positivamente.

4.4. Construção das Jogadas

No âmbito deste trabalho, uma jogada é uma cadeia de ações contíguas executadas pelo mesmo time, sem que a bola saia de campo, que sejam passes, cruzamentos, conduções, dribles e chutes. Dessa forma, uma sequência de ações que leve a um lateral ou escanteio é concluída no momento em que a bola sai de campo. Diferentemente de [McCarthy et al. 2023], foram incluídos não somente passes, mas também os outros tipos de ações mencionados anteriormente.

A construção do conjunto de todas jogadas segue os seguintes passos. Primeiramente, as ações de início e fim de cada uma das jogadas são identificadas. Em seguida, para cada jogada, os identificadores de cada uma das ações que compõem a jogada são armazenados, assim como as zonas correspondentes a cada uma das ações, os agrupamentos SoccerMix associados a cada uma das ações e o valor $P_{\text{scores}}(S_i, x)$ da última ação que faz parte da jogada. Além de armazenar essas informações, é definido um identificador para cada jogada, para que, ao analisar os resultados, seja possível recuperar as jogadas associadas aos padrões frequentes descobertos.

Um ponto a se notar, referente às sequências de agrupamentos SoccerMix associados a cada uma das ações que fazem parte da jogada, é que uma única ação pode estar associada a mais de um agrupamento, já que o método utilizado é de agrupamento suave. São considerados os agrupamentos de maior probabilidade, para cada ação, de modo que a sua soma seja pelo menos 90%.

Para metrificar cada jogada é escolhido o $P_{\text{scores}}(S_i, x)$ da última ação que compõe aquela jogada, e depois fazemos uma média desses $P_{\text{scores}}(S_i, x)$ para todas as vezes que aquela jogada ocorre. O uso do Pscores em comparação com o xG traz a vantagem de fornecer uma informação mais abrangente e contextualizada. Enquanto o xG calcula apenas a probabilidade de um gol ser marcado a partir de um chute específico, o VAEP quantifica todas as ações realizadas pelos jogadores, valorizando o impacto que cada uma delas tem na chance de um time marcar ou sofrer um gol. Isso permite a avaliação da sequência de ações que compõem uma jogada, em vez de usar somente a ação final da jogada.

O resultado das ações de enriquecimento é um arquivo que contem uma jogada enriquecida por linha, contendo as seguintes informações:

(i) identificador único da jogada;
(ii) time;
(iii) dados do jogo;
(iv) tempo de início e de fim da jogada;
(v) outros dados da jogada;
(vi) cadeia de identificadores de eventos que compoe a jogada;
(vii) VAEP da jogada;
(viii) xG da jogada;
(ix) cadeia de ações mapeadas para posições no grid; e
(x) cadeia de ações mapeadas para grupos do Soccermix.

A Tabela 1 apresenta algumas estatísticas dos vários times cuja estratégia foi caracterizada nesse trabalho. Em particular, é interessante notar o número relativamente pequeno de jogadas em relação a ações consideradas de cada time, sendo o comprimento médio das jogadas em torno de 20. Na próxima seção são discutidas a mineração de estratégias como sequências e suas medidas de interesse como critérios.

4.5. Mineração de Jogadas Típicas

Um dos desafios da caracterização de estratégias de futebol é identificar as estratégias típicas de cada time, representada por jogadas ou partes de jogada relevantes sob um ou mais critérios. Este trabalho adota sequências frequentes como um sumário da estratégia dos times. A mineração de sequências frequentes identifica todos os padrões sequenciais, ou seja, sub-sequências cuja frequência de ocorrência na base de dados de entrada é igual ou maior que um limiar pré-definido, chamado suporte. O suporte é um parâmetro fundamental para evitar a explosão combinatória do problema de mineração de sequências e permite descartar exceções e sequências acidentais.

Formalmente, um padrão sequencial é também uma sequência, $S = X_1, X_2, \ldots X_k$, onde cada X_i é um conjunto não vazio de ações. S é uma subsequência de uma jogada (sequência de ações) $j = Y_1, Y_2, \ldots Y_m$, onde cada Y_i é um conjunto não vazio de ações, se e somente se, existem inteiros $1 \leq i_1 < i_2 \ldots < i_k \leq m | X_1 \subseteq Y_{i_1}, X_2 \subseteq Y_{i_2}, \ldots X_k \subseteq Y_{i_k}$. Como mencionado, o número de jogadas onde S aparece é denominado $suporte(S)$. O algoritmo PrefixSpan[Pei et al. 2004] foi utilizado por ser considerado um dos mais eficientes para mineração de sequências. A estratégia então é minerar sequências frequentes a partir dessas cadeias. Em termos dos critérios, o suporte é a medida básica de frequência mínima e foi definido como 2%, ou seja, jogadas típicas

Time	Ações	Jogadas	Jogadas Típicas	Freq. Média	Imp. Média	PScore Médio	VAEP Médio
Arsenal	30521	1502	44290	44.9013	1.4128	0.0341	0.0317
Chelsea	28122	1484	13955	45.7976	1.4265	0.0207	0.0292
Manchester United	27063	1394	14860	41.9729	1.6323	0.0315	0.0365
Liverpool	30055	1527	21041	45.4093	1.5592	0.0428	0.0326
Newcastle	20882	976	2153	32.6795	1.7244	0.0283	0.0269
Southampton	24311	1081	8354	33.7572	1.5489	0.0497	0.0355
Everton	21875	867	3089	29.3729	1.4938	0.0239	0.0273
Tottenham	29077	1521	14586	46.5222	1.4412	0.0347	0.0295
Manchester City	35342	1751	107802	49.8424	1.5304	0.0206	0.0271
West Bromwich	19993	948	2154	31.7530	1.5621	0.0572	0.0297
Crystal Palace	21553	1145	2618	38.0672	1.5495	0.0400	0.0333
Leicester City	22157	1026	2179	34.4172	1.6309	0.0417	0.0336
West Ham	20959	908	2295	30.8009	1.7768	0.0243	0.0339
Stoke City	19436	864	1287	31.3054	1.5072	0.0142	0.0325
Watford	22697	1116	2407	37.7503	1.5704	0.0415	0.0280
Burnley	20048	940	1501	32.3404	1.5958	0.0403	0.0290
Brighton	21340	933	3327	30.7433	1.5361	0.0190	0.0348
Bournemouth	23414	1119	4270	36.7283	1.5306	0.0163	0.0252
Huddersfield	21901	919	2824	30.7624	1.6060	0.0295	0.0313
Swansea City	22155	844	4603	26.8644	1.6125	0.0338	0.0317

Tabela 1. Estatísticas sobre as ações, jogadas, jogadas típicas e medidas de interesse das sequências

tem que ocorrer em pelo menos 2% das jogadas analisadas. A importância é medida pela métrica de interesse aplicada a sequências. Seja uma sequência s, cujo tamanho em termos de ações é $|s|$ e tem a seguinte composição: $X_1 \rightarrow X_2 \rightarrow \ldots X_{|S|}$. A sua importância pode ser calculada conforme a Equação 4:

$$Imp(s) = \frac{\sum_{i=1}^{|S|-1} \frac{sup(S)}{sup(X_1 \ldots X_i) \times sup(X_{i+1} \ldots X_{|S|})}}{|S| - 1} \tag{4}$$

Finalmente, a efetividade é medida por duas métricas já discutidas: (i) PScore (Eq. 1) e (ii) VAEP (Eq. 3). A Tabela 1 apresenta os valores médios de suporte, importância, PScore e VAEP médios de cada um dos 20 times analisados. É perceptível que os valores variam bastante e não são correlacionados positivamente, ou seja, há frequentemente um compromisso entre os vários critérios, perceptível mesmo considerando as suas médias.

4.6. Identificação de Jogadas Típicas Frequentes, Importantes e Efetivas

Um desafio da caracterização alvo é que há um compromisso entre as várias medidas de avaliação dos padrões minerados. Por exemplo, a frequência e a importância podem ser inversamente correlacionados, assim como jogadas que são realmente efetivas. Uma forma de lidar com os compromissos entre várias métricas é calcular a chamada fronteira de Pareto [Godfrey et al. 2007]. Antes de mapear o conceito de fronteira de Pareto para jogadas típicas, é importante definir o conceito de dominância. Sejam duas sequências

frequentes s_1 e s_2 e dois critérios a e b, define-se que s_1 domina s_2 se $a(s_1) \geq a(s_2)$ e $b(s_1) \geq b(s_2)$. A definição básica da fronteira de Pareto é que ela consiste exatamente naquelas alternativas que não são dominadas por nenhuma outra alternativa. Um resultado ótimo de Pareto é aquele em que nenhuma jogada pode ficar melhor sem piorar a situação de outra jogada.

Esse trabalho considera fronteiras de Pareto tridimensionais, que buscam as jogadas típicas que representam um compromisso entre os três critérios discutidos na Seção 3. Em particular, os resultados apresentados na próxima seção se baseiam em duas fronteiras de Pareto. A primeira fronteira, denominada $\mathcal{FIP}$, considera frequência, importância e PScore, enquanto a segunda, denominada $\mathcal{TIP}$, considera tamanho da sequência, importância e PScore. A motivação para avaliar essa segunda configuração de fronteira é que considerar o suporte tende a priorizar jogadas típicas com menos ações, enquanto o tamanho da sequência tende a priorizar jogadas típicas com mais ações. Como vai ser mostrado, essas configurações são complementares em relação à informação provida. Embora tenham sido avaliadas outras configurações para fronteira de Pareto, as duas escolhidas se mostraram mais informativas.

5. Resultados

Nesta seção são apresentados alguns estudos de caso que exemplificam resultados obtidos com a aplicação do modelo proposto neste trabalho. Os resultados que serão apresentados ilustram que a metodologia proposta permite caracterizar estratégias dos diferentes times.

Comparação *SoccerMix* - Grade: Considerando as fronteiras de Pareto $\mathcal{FIP}$ das jogadas típicas a partir de agrupamentos gerados pelo SoccerMix e as zonas estáticas pré estabelecidas pela grade, pode-se perceber uma grande diferença entre as caracterizações das jogadas. No caso do Leicester City (Figura 2), apesar de ambas as fronteiras apresentarem semelhanças visuais, é nitido o ganho de informação com o uso do *SoccerMix*, tendo em vista a abstração das ações em agrupamentos mais representativos do que as zonas. Enquanto a grade é um método estático, combinando todas as ações ocorridas em uma zona, o modelo *SoccerMix* assinala probabilidades de cada ação pertencer a um agrupamento levando em consideração, além de sua localização, o seu tipo, com agrupamentos de cruzamento e drible em posições ofensivas pelos lados, realçando detalhes da estratégia do time caracterizado.

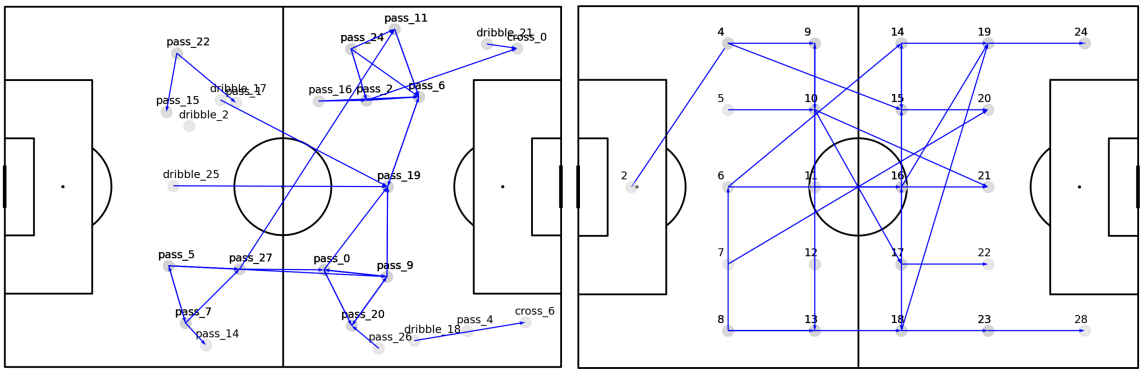

Figura 2. Pareto $\mathcal{FIP}$: SoccerMix e Grade - Leicester City

Representatividade das jogadas típicas: Uma propriedade fundamental das jogadas típicas para fins de caracterização é a sua representatividade, ou seja, o quanto ela sintetiza as jogadas cuja mineração resultaram nela. Para melhor visualizar essa representatividade, pode-se apresentar tanto jogadas típicas quanto as jogadas a ela associadas. Dessa forma, é possível extrair padrões de como as jogadas de um time são representadas pelas localizações prototípicas do SoccerMix. Na Figura 3, são apresentadas duas jogadas típicas (em vermelho), uma do Tottenham (esquerda) e outra do Huddersfield (direita) e as jogadas que geraram (azul) essa jogada típica. No caso do Tottenham, há uma jogada típica, de alto valor P_{scores}, composta por passes, que se inicia no próprio campo de defesa e avança consideravelmente no campo. Percebe-se que uma quantidade significativa das jogadas associadas a essa jogada típica resultam em alcançar a área adversária, demonstrando que possivelmente são jogadas elaboradas, que se iniciam na defesa e chegam até a fase final de ataque. A jogada típica de alto valor P_{scores}, destacada do Huddersfield, time de qualidade consideravelmente abaixo do Tottenham, envolve não somente passes, como uma localização de cruzamento também. Além disso, nota-se que a maioria das jogadas associadas a essa jogada típica ocorrem pelo lado direito do time, frequentemente tendo como última ação o cruzamento "cross_8".

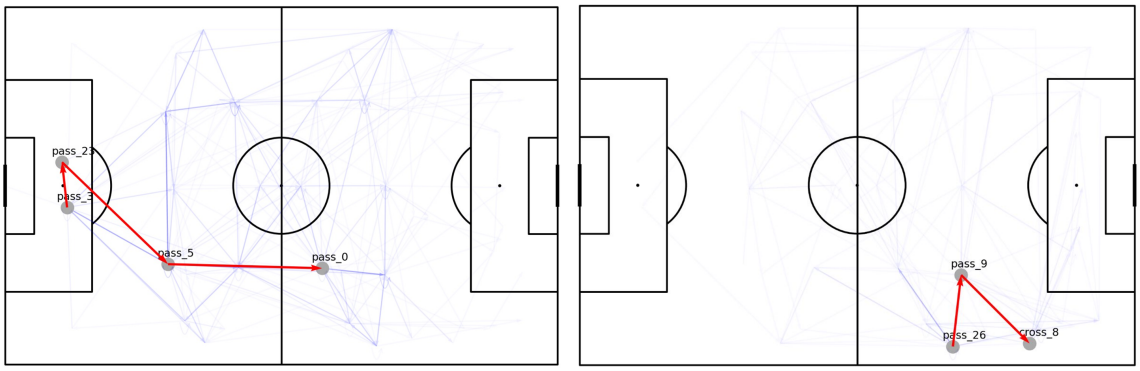

Figura 3. Jogadas típicas $\mathcal{FIP}$: Tottenham e Huddersfield

Pareto como sumário estratégico: A fronteira de Pareto condensa várias jogadas típicas, sumarizando a estratégia do time. Para ilustrar essa possibilidade, a Figura 4 apresenta as fronteiras de Pareto $\mathcal{FIP}$ de Manchester City e Liverpool, e a Figura 5 apresenta os mapas de calor das ações dos mesmos times. Pode-se perceber com clareza duas linhas de passes e jogadas típicas do Manchester City, com significativas participações nas alas do campo. Tanto a linha defensiva quanto ofensiva de construção de jogadas nos mostram a força dos flancos e a grande quantidade de jogadas, enquanto na região do meio campo não temos tanto volume. Isso representa bem o estilo do técnico do time, Pep Guardiola, com uma linha de 5 próxima a área do adversário, sempre com a pressão rápida e presença nas laterais do campo. Já a mesma visualização do Liverpool mostra a centralização do jogo desse time, favorecida pelo meio campo forte e pela construção de jogadas pelo meio do campo. Além disso, esse estilo de jogo é favorecido pela centralização de ambos os pontas do time, Mané e Salah, que não se isolam muito nas laterais, e do centroavante Roberto Firmino, conhecido por ser também um grande criador de jogadas e se aproximar dos meio campistas. Dessa forma, percebe-se que a fronteira de Pareto gerada a partir do

SoccerMix conseguiu transmitir bem algumas das características dos dois times com mais gols marcados no campeonato.

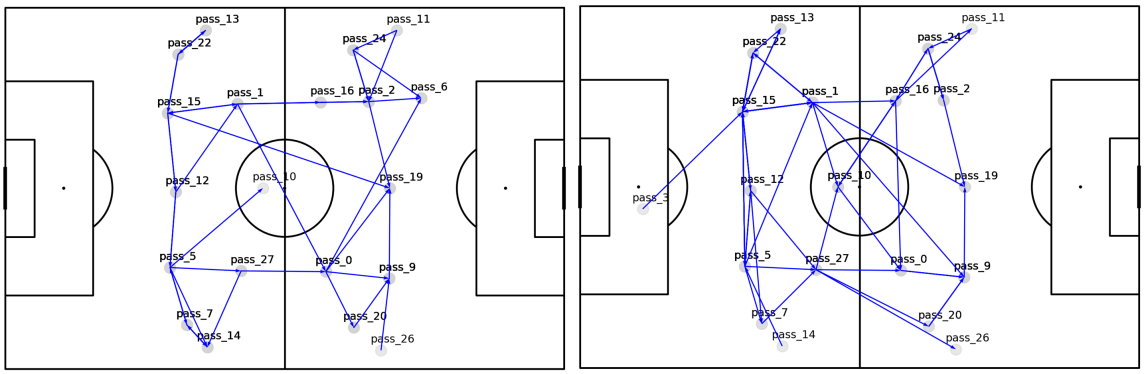

Figura 4. Pareto $\mathcal{FIP}$: SoccerMix - Manchester City e Liverpool

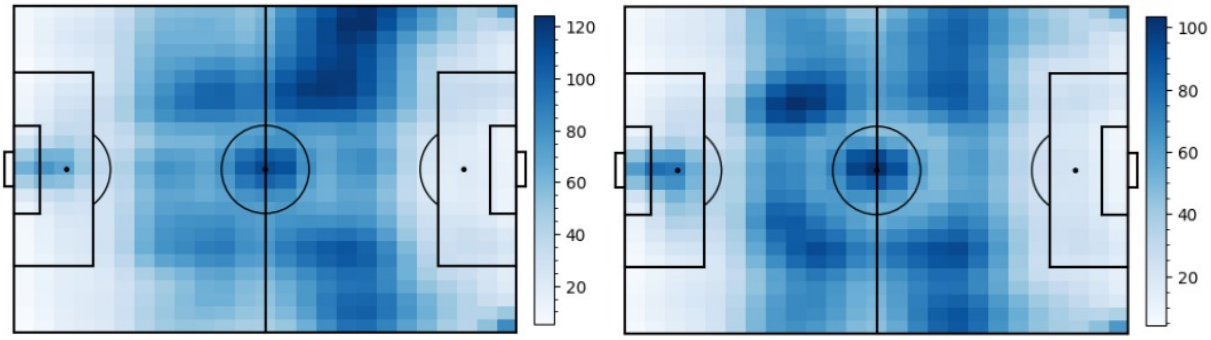

Figura 5. Mapas de Calor de Ações - Manchester City e Liverpool

Complementaridade entre as fronteiras de Pareto $\mathcal{FIP}$ e $\mathcal{TIP}$: Como mencionado, as fronteiras de Pareto $\mathcal{FIP}$ e $\mathcal{TIP}$ se diferenciam pela primeira dimensão considerada, que é a frequência das jogadas típicas e o tamanho das mesmas, respectivamente. Um bom exemplo dessa diferença pode ser percebida considerando o time Tottenham (Figura 6), onde percebe-se que o Pareto $\mathcal{TIP}$, que prioriza o tamanho das jogadas, identifica um conjunto de jogadas que não reflete a terceira colocação obtida naquele campeonato pela equipe e nem seu grande número de gols marcados. Por outro lado, a fronteira de Pareto $\mathcal{FIP}$, que prioriza a frequência, apresenta mais sequências ofensivas e uma rede defensiva de construção bem mais elaborada. Isso ilustra a objetividade e reatividade do time do Tottenham, que ataca com perigo pelos lados e em jogadas objetivas e rápidas, com poucas jogadas com muitas ações e com mais jogadas curtas e frequentes, sendo um time objetivo e de avanço agressivo para o ataque.

6. Conclusão e Trabalhos Futuros

Este trabalho apresentou uma nova proposta para caracterizar estratégias de futebol a partir de eventos como passes, dribles e chutes a gol. Foram utilizadas duas modelagens desses eventos como ações e as jogadas típicas foram identificadas aplicando um algoritmo de mineração de sequências. Cada jogada típica é avaliada considerando quatro critérios e foram avaliados os compromissos entre esses critérios através de fronteiras de Pareto.

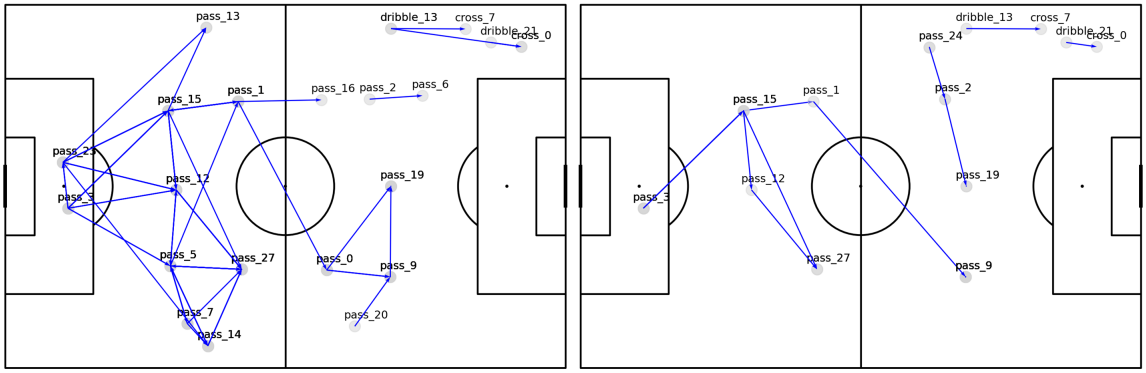

Figura 6. Pareto $\mathcal{FIP}$ SoccerMix - Tottenham

Utilizando dados reais, foi possível avaliar diferentes aplicações da metodologia proposta, na qual a escolha do SoccerMix se mostrou valiosa em relação à grade para fins de agrupamento espacial das ações, oferecendo uma representação mais informativa e detalhada das estratégias dos times. Ao combinar os agrupamentos do SoccerMix com as fronteiras, foi possível identificar as jogadas típicas dos times, caracterizando suas estratégias ofensivas. Além disso, por meio da utilização de diferentes critérios na construção das fronteiras de Pareto, foi possível entender melhor as estratégias adotadas.

Como trabalhos futuros, existe uma gama de possibilidades de avaliação dos resultados gerados, que podem ser usadas para responder diferentes questões sobre perfil de jogadas e estratégias de jogo de futebol. Além disso, pretende-se estender o trabalho para incorporar novas métricas de avaliação no modelo proposto.

Agradecimentos

Este trabalho foi parcialmente financiado por CNPq, CAPES, FAPEMIG, MASWEB e IAIA (INCT para IA). Agradecemos Pedro Picchioni, *Head of Scouting Analytics*, e Rodrigo Picchioni, *Head of Football Analytics*, ambos do Clube Atlético Mineiro, pelo apoio e discussões inspiradoras deste trabalho.

Referências

[Davis et al. 2022] Davis, J., Bransen, L., Devos, L., Meert, W., Robberechts, P., Van Haaren, J., and Van Roy, M. (2022). Evaluating sports analytics models: Challenges, approaches, and lessons learned. In *AI Evaluation Beyond Metrics Workshop at IJCAI 2022*, volume 3169, pages 1–11. CEUR Workshop Proceedings.

[Decroos et al. 2019] Decroos, T., Bransen, L., Van Haaren, J., and Davis, J. (2019). Actions speak louder than goals: Valuing player actions in soccer. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1851–1861.

[Decroos et al. 2020] Decroos, T., Roy, M. V., and Davis, J. (2020). Soccermix: Representing soccer actions with mixture models. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 459–474. Springer.

[Decroos et al. 2018] Decroos, T., Van Haaren, J., and Davis, J. (2018). Automatic discovery of tactics in spatio-temporal soccer match data. In *Proceedings of the 24th ACM*

SIGKDD International Conference on Knowledge Discovery & Data Mining, pages 223–232.

[Ensum et al. 2004] Ensum, J., Pollard, R., and Taylor, S. (2004). Applications of logistic regression to shots at goal in association football: Calculation of shot probabilities, quantification of factors and player/team. *Journal of Sports Sciences*, 22(6):500–520.

[Fernández et al. 2021] Fernández, J., Bornn, L., and Cervone, D. (2021). A framework for the fine-grained evaluation of the instantaneous expected value of soccer possessions. *Machine Learning*, 110(6):1389–1427.

[Godfrey et al. 2007] Godfrey, P., Shipley, R., and Gryz, J. (2007). Algorithms and analyses for maximal vector computation. *The VLDB Journal*, 16(1):5–28.

[Green 2012] Green, S. (2012). Assessing the performance of premier league goalscorers. *OptaPro Blog*.

[Harper 2021] Harper, J. (2021). Data experts are becoming football's best signings. *BBC News*.

[Malqui et al. 2019] Malqui, J. L. S., Romero, N. M. L., Garcia, R., Alemdar, H., and Comba, J. L. (2019). How do soccer teams coordinate consecutive passes? a visual analytics system for analysing the complexity of passing sequences using soccer flow motifs. *Computers & Graphics*, 84:122–133.

[McCarthy et al. 2023] McCarthy, C., Tampakis, P., Chiarandini, M., Randers, M. B., Jänicke, S., and Zimek, A. (2023). Analyzing passing sequences for the prediction of goal-scoring opportunities. In *Machine Learning and Data Mining for Sports Analytics: 9th International Workshop, MLSA 2022, Grenoble, France, September 19, 2022, Revised Selected Papers*, pages 27–40. Springer.

[Pappalardo et al. 2019] Pappalardo, L., Cintia, P., Rossi, A., Massucco, E., Ferragina, P., Pedreschi, D., and Giannotti, F. (2019). A public data set of spatio-temporal match events in soccer competitions. *Scientific data*, 6(1):236.

[Pei et al. 2004] Pei, J., Han, J., Mortazavi-Asl, B., Wang, J., Pinto, H., Chen, Q., Dayal, U., and Hsu, M. (2004). Mining sequential patterns by pattern-growth: The prefixspan approach. *IEEE Trans. Knowl. Data Eng.*, 16(11):1424–1440.