Capturing Team Strategy in European Football via Process Mining and Visual Analytics

Marc Garnica-Caparrós, Dilyar Iskan, and Daniel Memmert

Abstract—The use of data and machine learning has significantly contributed to the analysis of sports performance, particularly in invasion sports like European football. Event data, which provides detailed information on every action taken during a game, became one of the most widely used sources of information. The study of team tactics, which is essential in understanding analytical outcomes, involves the examination of a team's playing style, characterized by its unique playing patterns, processes, and actors. This paper proposes a process-aware analysis of event data to discover team playing strategy. The approach employs Process Discovery techniques to retrieve and analyze patterns of play out of event traces. A purpose-driven methodology reduces the variance and structurelessness in the traces, highlights frequent patterns of play independently of their success, and facilitates the identification of accurate models of team strategy. Moreover, despite the lack of ground truth to validate the findings, the models are evaluated in terms of fitness against the observed behavior and generalization toward new data. The proposed methodology is evaluated in one season of the English Premier League, where teams' strategies to perform a certain task are identified and visualized, and their resilience to these strategies is measured throughout the season.

Index Terms—Sports Analytics, Event Data, Team Tactics, Process Discovery, Visual Analytics.

1 Introduction

Over the last two decades, the sports industry has incorporated data analytical methods in their daily methodologies as a competitive advantage. The systematic usage of data-driven methods enabled sports teams to retrieve objective performance analysis, improve understanding of the sports discipline and refine knowledge extraction. In invasion sports such as European football (soccer), the exponential growth of data volume and detail motivated the introduction of more sophisticated machine learning models to handle the complexity of the analysis [1]. Such data empower a completely new reconstruction of the sports game beyond the traditional performance indicators and match sheet data [2]. Among these data sources, event data rose as one of the most widely used sources of information.

Event data reports at high granularity all actions occurring with the ball during a sports game, including key attributes such as the players involved or the precise location of the event in the field. In practice and research, this data source is often set side by side with optical tracking data, a data source including all the players and the ball position at high frequency during the game [3]. Despite the disadvantages of using event data-driven analysis, often related to missing crucial off-ball contextual information, event data has become predominant and more adopted than optical tracking data analysis in modern sports clubs and stakeholders because of its higher availability and finer descriptions of the events. Being both data sources available, the ideal scenario would consist of synchronizing the data streams to gain the best of both data contexts. Event data providing the detail description of the events occurring and tracking data reporting the precise positions of all the 22 players involved and the ball. However, these would introduce inaccuracies since a proper synchronization method for any arbitrary event has yet to be reported, mainly because of the two data sources coming from different collection technologies and system clock discrepancies [4], [5].

The usage of data and machine learning in team sports can be divided into three areas: player recruitment, match preparation, and health management [6]. Among these challenges, the concept of team tactical strategy, sometimes also called team style [7], is essential in understanding the analytical outcomes. For instance, when assessing a potential new player for a team, it is desirable to know the systems or mechanisms that the team intended to perform and how this player contributed to this team's strategy. On some other occasions, when approaching a certain opponent, some of the common questions the analyst team faces at a high level are 'how are they usually trying to generate chances?', 'what is their ball recovery strategy?'. However, for obvious reasons, this team strategy is kept private as it belongs to the coaching staff and is highly valuable for the team's success.

Technically, event data is interpreted as a time-ordered sequence of events. The proper modeling of this sequence of events can be used to objectively measure the impact of each action performed in the field, for instance, with probabilistic classification approaches toward reaching a certain reward (i.e., scoring opportunity). Recently in [8], the

- *M. Garnica-Caparrós and D. Memmert are with the Institute of Exercise Science and Computer Science in Sports, German Sports University Cologne, Am SportPark Müngersdorf 6, Cologne, 50933, Germany*
- *D. Iksan is with the RWTH Aachen, Ahornstr. 55, Aachen, 52056, Germany.*

For correspondence with regard to this manuscript, please contact the corresponding author Marc Garnica-Caparrós.
This research was supported by the German Sport University of Cologne grant HIFF 1 2022

authors proposed a method based on estimating scoring and conceding probabilities at any given moment of a football game. However, these approaches are usually centered on probabilistic predictions and game-state measuring but do not directly focus on identifying team playing strategies. Moreover, these approaches usually focus on success rather than systematic play patterns. This paper explores a different approach to discovering and modeling team tactics of play by using event data as a trace of such a strategy. Indeed, assuming teams have a set of principles of play defined before a match that describes how the team should unfold in certain game situations, event data can be considered as the data footprint or trace of such a system.

As denoted in [7], team playing style can be defined as "the characteristic playing pattern demonstrated by a team during games." Despite the media presenting football tactics at a high level on how a team esthetically looks or simplifying it to aggregated metrics[1], tactics in football are divided into various levels and categories. Depending on the game situation, the team undergoes a defined process involving actions and actors (i.e., the players). For instance, a team develops different processes when building the attack or recovering the ball [9]. Additionally, an attacking process can pursue different objectives or endeavors, such as reaching a certain zone in the field, conserving a dangerous possession, or directly attacking the opponent's penalty box at a fast tempo. All these processes are part of the team's guidelines, automatisms, and voluntarily executed actions from the players that, if repeated frequently, construct the team strategy.

The end goal of this paper is to present a process-aware analysis of event data to discover team playing strategy. We use Process Discovery [10] techniques to retrieve and analyze the inherent patterns of play out of event traces extracted from event data. We present a purpose-driven methodology to reduce the variance and structurelessness in the traces, highlight frequent patterns of play independently of their success, and facilitate the control flow identification to produce the most accurate models of team strategy. Hence, we contribute to defining a knowledge discovery pipeline to manage and analyze the large-scale availability of event data for team strategy identification. We also demonstrate how the models could be integrated into football-specific visualizations to provide a novel and better understanding of a team's execution during a game. The methodology is evaluated in the 2017/2018 season of the English Premier League. Teams' strategies to perform a particular task are described and visualized. Additionally, we measure teams' resilience to their strategies throughout the season.

2 Related work

To contextualize the rest of the paper, this section provides the background on Process Mining methodologies for unstructured process analysis and the main contributions regarding the motivation use case of identifying playing strategy in football.

1. https://theanalyst.com/eu/2022/01/premier-league-how-does-each-team-play/, accessed 11/11/2022

2.1 Process mining unstructured systems

Process Mining (PM) is a semiautomatic evidence-based methodology to discover, monitor, evaluate and improve processes [11]. It untangles the differences between the event logs of a behavior or system and the actual processes. PM combines traditional data-driven models sourcing from Business Project Management research with modern data mining and machine learning techniques to identify recurrent patterns in historical event data and shape the inherent process models. Process discovery allows an automated definition of a behavioral process from the collected event data [12]. Additionally, conformance checking techniques provide methods to measure the deviations between the theoretical processes (i.e., plan or strategy) with the processes seen in reality (i.e., execution) [13]. These methods leverage the variants and deviations that might occur between the occurring events and the process model underneath [14].

The process-aware analysis assumes the collected event data results from a dynamic behavior process. Roles and actors interact towards a common objective or function in time and execute certain patterns, orders, and flows. Other analysis techniques, such as episode mining [15], also perform pattern-matching tasks on sequences but do not consider the end-to-end process and the set of actors. Processes can be structured or unstructured. Examples of structured processes can be found in event logs sourcing from digital systems such as websites or transaction-based information systems. In these processes, cases are highly stable, and deviations from the theoretical process are scarce. Process models are elaborated as part of the system's design phase, and PM is an efficient evaluation tool to ensure the transactions are being executed as expected. However, as denoted by [14], PM popularity in industry and research has increased in recent years from various applications and domains. Processes no longer must have a predefined structure, and they combine digital elements with physical flows.

Therefore, the model discovering algorithm cannot assume that all possible behavior is reported in the event logs. Integrating such natural processes in PM analysis can bring invaluable knowledge and significant benefits, but it is also more challenging. [16] present a methodology to handle unstructured processes by combining algorithms, narrowing the event logs cases, and visual analytics. In an environment where the inherent process is unknown, and the event logs contain a high heterogeneity, innovative data-driven methods effectively exploit the collected event data by leveraging noise and low-frequent behavior as the Heuristic Miner [17]. [18] introduce a frequency-based approach, the Fuzzy Miner, producing a process map representing the precedence relationships in the sequences. Model overfitting or underfitting can be tuned by modifying the frequency threshold on nodes and paths to filter out infrequent behavior. The incompleteness of the observed data is tackled by [19] providing a modification of the Inductive Miner with probabilistic behavioral relations that allow for a more accurate approximation to the original system. We refer to [20] for a comparison and detailed explanation of the existing algorithms for sequence mining and process-aware discovery.

Sequential pattern mining aims to find the frequent subsequences in a sequence data set. Thus, they assume there is only a single correct answer. Although process discovery algorithms use support-based methods from sequential pattern mining, they differ in the purpose of discovery and output. Process discovery algorithms utilize frequent sequence patterns to define real behavior captured in event data. In such cases, the same event data set can be abstracted to a different view depending on the analysis requirements, environment, and data collection characteristics. The visual representation of the discovered models also plays an essential role in the discovery training and evaluation. Furthermore, a process discovery model's output is represented visually and conceptually by means of a process modeling language [21]. These languages can be notation-based graphical representations of workflows like BPMN [22] or mathematical models representing discrete variables, states, and transitions such as Petri Nets [23] and extensions like the YAWL language [24].

2.2 Team strategy data-driven identification in football

Event data is the time-ordered collection of all actions occurring in a football game [25]. The event information consists of systematic information such as the timestamp when the event occurred, the key actor of the event (e.g., the player acting), the team, the spatiotemporal features of the event (i.e., x and y coordinates of the event in the playing field), and the outcome of the event (e.g., the success of a pass). Other attributes that event data might contain include but are not limited to the part of the body used to perform the action (e.g., left foot), subcategorization of the event (e.g., disseminating between diagonal passes and vertical passes), or the difficulty of the event (e.g., postprocessing information on the number of defenders in front, or the position of the goalkeeper in the event of a shot). The advances in data collection technologies, optical tracking data, and data labeling enable the addition of significant attributes to each event in almost real time[2].

The authors in [7] identify player and ball movements as highly important in determining a team's style in terms of time and space. Strictly excluding any information from the events occurring in a football game, tracking data still conveys much information on how the team is playing. For instance, team formations can be identified using tracking data [26]. Football team formations describe the roles of each player on the field and are highly coupled with team tactics and strategy [27]. Team tactics are expected to include guidelines and rules for every game phase. In [28], tracking data analysis can also identify different game phases; moreover, team tactics are not approached holistically but in a narrowed environment with clear team objectives, for instance, when performing counter-pressing. The granularity and frequency of tracking data can also retrieve collective metrics such as team centroids or space control measures relevant for identifying how a team attacks [29]. Tracking data also can help shed light on the collective behavior of teams depending on their defensive strategies [30]. However, using event data, the players and ball positions are available only when related to an on-ball event, and the analysis only accounts for the scarce positions reported at event time. Nevertheless, several contributions obtained highly valuable tactics and insights from event data.

The first basic approach computed count-based aggregated metrics from event data that extended the match sheet information with new metrics accounting by time and space. These statistics at the team or player level provide some insights into the style of teams [9], [31], being able to associate "direct" or "possessional" plays by the number of passes per attack, passing speed rate and pass direction (i.e., direct and sideway passes). Most advanced approaches acknowledge the sequentiality of the event data source. On the one hand, methods like VAEP [8] use probabilistic classifiers to value each action or a subset of actions occurring in the game. Once every action has been assigned a value, teams or players can be categorized by how much they rely on their contribution to success. For instance, team A increase in winning probability can come from their passing ability in the midfield, while team B relies on long individual dribbles.

On the other hand, the sequentiality of event data can be exploited by finding spatiotemporal patterns. In these cases, finding frequent sequences of events within football possessions is challenging because possessions vary greatly in length, actions performed, players involved, and location on the field. [32] extend the concept of network motifs applied to passing sequences in football [33] to provide a tactical and statistical analysis of passing behavior in football. Teams and players are clustered in different groups based on their most frequent motifs and difference arose between wide and narrow styles of plays concerning their pass sequences. However, these motifs do not include any spatiotemporal information about the events. In [32] contribution and similar approaches, the time between passes is restricted; however, in a process-aware analysis, the time between activities is taken into account for the process discovery [12]. In a similar approach to this paper, [34] applies clustering techniques and a success score metric to discover frequent sequential patterns at the team level. In recent research, PM has been preliminary explored to provide a process-aware analysis of team strategy playing in ball possession phases and attacking sequences, where on-ball events are more frequent[3]. [35] show the potential of end-to-end process analysis of football event data focusing on a small sample of games and describing the new insights and visual analytics a methodology like PM can provide. Process discovery techniques are able to retrieve information on frequent patterns and players' collaboration.

3 Constructing Team Traces from Event Data

A football match comprises several situations differing in context, players involved, location on the field, and moments of the match. The variability of events that occur

2. 360 Data. The Industry's Most Detailed Soccer Data, https://statsbomb.com/what-we-do/soccer-data/360-2/, accessed 11/11/2022

3. Process Mining Meets Football! How Does a Football Team Possess The Ball On The Pitch. https://fluxicon.com/blog/2019/10/process-mining-meets-football-how-does-a-football-team-possess-the-ball-on-the-pitch/, accessed 08/11/2022

in a game, together with the heterogeneity of goal-setting occasions that players and teams phase at different moments of the game, requires a systematic approach to process event data and untangle team-style models. For instance, teams might arrange their players in a certain formation and perform a set of tactics to reach the opponent's goal when having full control of the ball. At the same time, in other cases, they might focus on maintaining their ball and restructuring their attack. Depending on these circumstances, the team might display different behavior patterns, and it is crucial to capture and analyze these patterns in concordance with the team's objectives.

3.1 Team in-play purpose-driven sequences

The proposed methodology for processing event data, identifying purpose-driven team sequences, and discovering team patterns is displayed in Fig. 1. Initially, the raw event data is ingested. Event data might come in slightly different formats depending on the data provider. However, most of the syntax and semantics of the data will include a chain of events performed by any of the two teams. The second step involves dividing these large sequences into subsequences. Thus, the raw event sequence can be divided into sequences where each team has the ball, usually referred to as possessions. A possession begins when a team gains the ball and ends when the match is interrupted (i.e., ball out of bounds, end of a period, or foul), the team in possession scores or the other team regains the possession of the ball.

In order to gain the maximum transparency and interpretability of the team strategy models, the next step involves defining specific analysis questions translated into team objectives in the field. Team possession traces are filtered to obtain all the possessions where the team acted with common characteristics, specifications, or outcomes. While knowing the ground truth of what the team aimed at a certain moment of the game is impossible, traces can be filtered by how they started (e.g., location in the field), the player who gained the ball, or the number of certain events. This purpose-driven filtering reduces the variability in the actions to benefit the discovery and interpretation of the models. Examples of analysis questions that could benefit from the creation of these team purpose traces could be but are not limited to: Traces that reach a certain zone in the field (e.g., the opponent penalty box, zone 14, offensive third, etc.), traces starting from the goalkeeper and end in a turnover in the team's own half, or traces that start with a recovery in the team's own penalty box and end in the offensive third in less than 15 seconds.

3.2 Field partitioning

Ultimately, one of the key features of the team purpose traces is the locations in the field the team is reaching to achieve their immediate goal. However, by using raw spatial data of the events, the model might miss important insights as raw locations can be noisy and contain a lot of small variations. In order to identify larger-scale patterns and trends in field usage, team purpose traces are also spatially aggregated using a grid-based partitioning method where the entire field is divided into a grid of cells, and the event location is assigned to the cell that falls within. Field partitioning allows the discovery of models to identify similar patterns even if the events' trajectories are not exactly the same. The choice of method for partitioning the field could be exchanged depending on the analytical use case [36], [37]. An example of a simple field partitioning with highlighted common cells between two event sequences is shown in Fig. 2. Overall, the choice of partitioning method will depend on the specific characteristics of the data, the purpose filtered developed, and the overall team tactics. Coaches are usually defining their field partitions with custom channels and zones. Other data-driven split criteria are available such as hierarchical partitioning or trace clustering. However, grid-based partitioning methods are often simpler and more computationally efficient.

4 Discovery and evaluation of team processes

4.1 Process mining team purpose traces

Once the team traces have been defined, the goal of the final steps of the methodology proposed is to configure the process discovery approach and interpret the resulting artifacts. Team traces are transformed into a compatible format for the Heuristic Miner algorithm. For each trace, several attributes are described accordingly to the conceptual model so that the discovery can be executed.

The **case ID** of a trace is the unique identifier assigned to a specific instance of a process. It groups all the events that belong to a single process instance. Thus, every team purpose trace is assigned a unique identifier that will allow the mining algorithm to group the events and treat the sequence as a trace. The **activity ID** is the identifier assigned to a specific task or activity in the trace. The aimed process is assumed to consist of different actions or steps identified by the activity ID. Every event on each team purpose trace must have a certain activity ID assigned. If we chose to identify the activity with the action type of the event (e.g., a pass or a shot), the process discovery would be sound. However, it will miss contextual information as any pass in the match would be treated as the same action for the mining algorithm. Therefore, the activity ID of an event is configured as the compound key between the action type and the assigned zone in the field after adding the field partitioning. So a step in the process is modeled as a certain type of action in a certain field zone. Additionally, other attributes are indicated to the mining algorithm, such as the event's resource (i.e., the player performing the event) and the timestamp of the event.

The discovery process ingests the traces of all the event logs and produces two artifacts as outputs, a Petri Net and a Heuristic Net. The generated traces contain critical challenges that differ from common process mining use cases, and they contain a high level of variance and noise that can make extracting behavior patterns difficult. We base our discovery approach on the Heuristic Miner (HM) [17], [38], where they use frequency information and dependencies to discard the noise and incompleteness of the logs. The miner algorithm constructs a dependency graph and a causal matrix accounting by a dependency threshold definition. We opted for a threshold dependency of 50%.

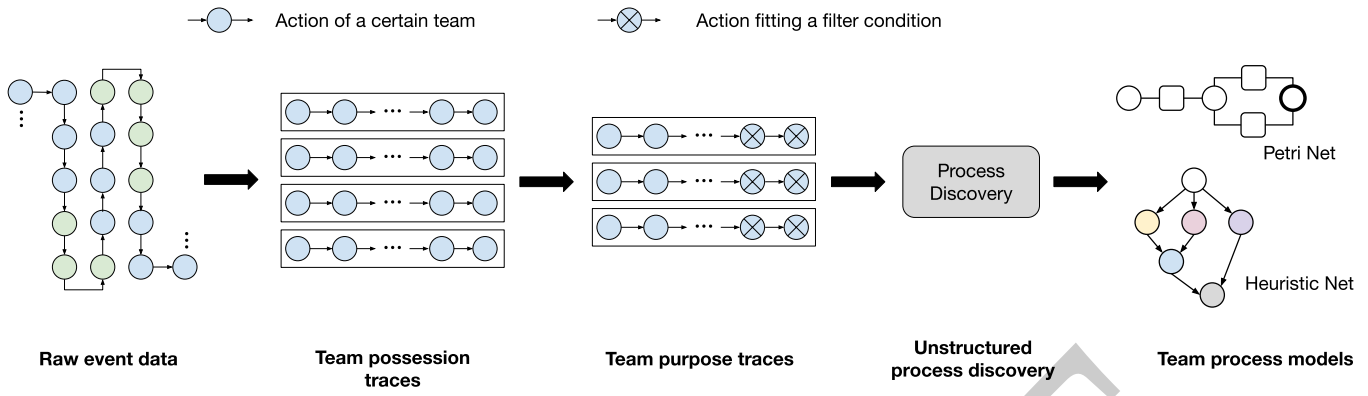

Fig. 1. Process Discovery methodology. Raw event data is mapped into team trace logs containing the sequences performed by each team. Certain criteria must be moduled to filter these traces to remove noise and construct the team purpose traces with similar sequences (similarity defined by arbitrary criteria such as type of events, location on the field, or players involved). Team purpose traces are fitted into a Heuristic miner where team behavior patterns are described employing a Petri net and a Heuristic Net.

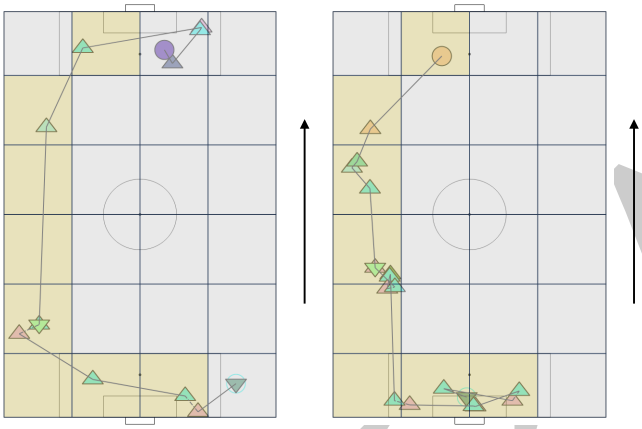

Fig. 2. Similar team traces starting on the defensive side of the field and being able to reach the opponent's penalty box. While the traces look similar, the raw location of the events might induce misleading differences. A simple field partitioning helps highlight the common zones the team used to build the sequence up and access the opponent penalty box.

Thus, the algorithm considers dependency values over 50% to ensure we are not capturing low frequent behavior but also allows the final model to be generalizable to new data. We used the implementation of the HM presented in the python package PM4PY [39]. All the other parameters of the discovery method were kept as default. In order to facilitate a closer examination of the generated models, we have included examples of Petri nets and Heuristic nets in the supplementary material. We believe that this supplementary material will be useful for readers who wish to explore the topic further or gain a deeper understanding of the process models presented in this article.

4.2 Evaluating the models: Fitness vs. Generalization

The evaluation of the extracted models is also subject to complexity as there is no objective ground truth documenting the strategy of football teams or the tactics deployed during certain moments of a game. Therefore, the models' validity is evaluated in two phases. First, the correctness of the model is evaluated in terms of the recall measure, usually referred to in PM as fitness. Model fitness refers to how much the generated model can execute the observed event logs. Model fitness can be easily computed by replaying the traces of the event log into the generated model and assigning a trace fit if it can be executed in the model. Last, the model's usefulness in describing a team strategy is measured by its generalization. Generalization tries to quantify how much a model can fit unseen behavior. We refer to [40], [41] for a detailed description and comparison of these three metrics. We use the generalization measure implemented in PM4PY and based on the work of [40].

4.3 Heuristic Maps: Translating process models to football

The process models are represented utilizing two different artifacts. A Petri net consists of places, transitions, and directed arcs between them. Places represent conditions or states, while transitions represent events or actions that can change the system's state. The arcs represent the flow of tokens, which represent resources or objects that move between places and trigger transitions. In reality, states might be concurrent executions from the event log, and transitions might refer to certain events or sets of events that trigger a state change in the system. Conversely, a Heuristic Net is a causal network involving actions or tasks as nodes. Arcs between these actions represent the dependency between them. These arcs are weighted with an indication of relative dependencies between nodes. Thus, this network structure infers dependencies between actions in the event logs and allows an understanding of how a certain overall goal is performed by dividing it into several steps and dependencies between them. Both model representations can be explored to understand the logic extracted by the discovery algorithms. However, often translating these models into practical information takes time and effort. Therefore, we also developed a domain-driven mapping of these logical structures into the so-called heuristic maps. Heuristic maps stem from the basic ideas presented in existing contributions, such as passing networks or passing flows. However,

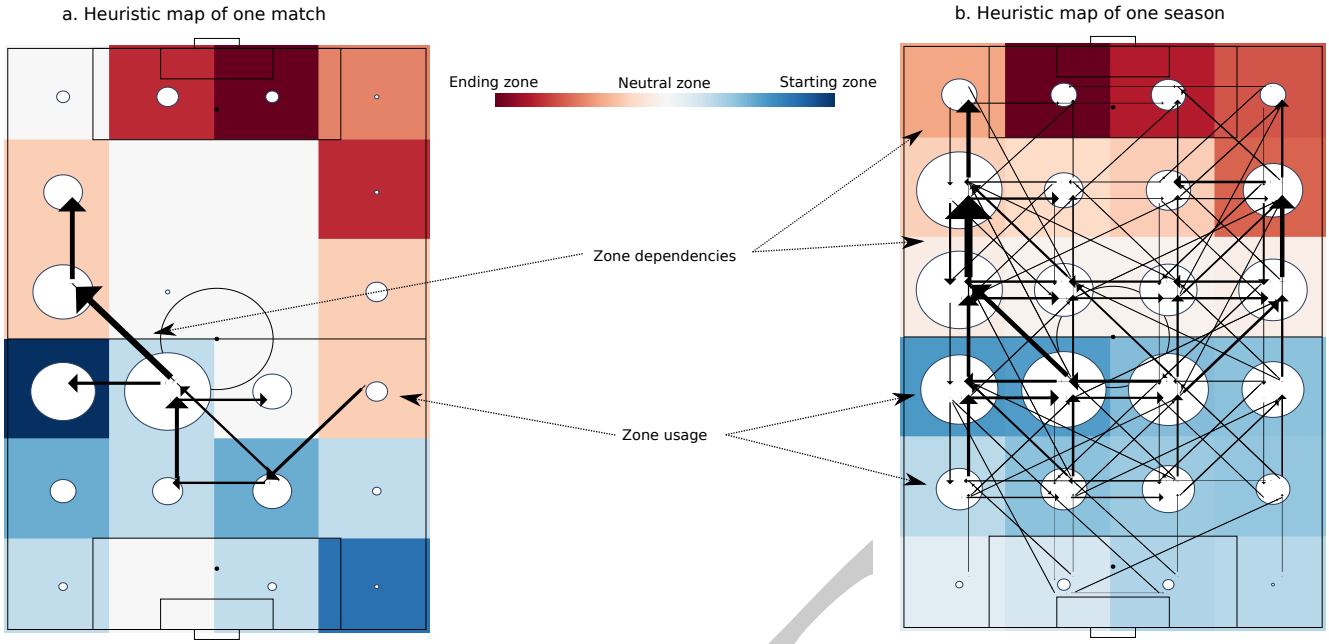

Fig. 3. Two examples of Heuristic maps with their components, a Heuristic Map of one match (a) and a Heuristic map of the same team for a whole season (b).

its content is based on the information represented for the causal network produced by the Heuristic Miner. A detail explanation of the proposed Heuristic maps visualization is presented in Fig. 3. Examples of heuristic maps can be seen in Section 5. To construct a Heuristic map from a Heuristic network, the following steps are followed.

- **Start and end zone**. Each zone of the field is assigned a standardized weight from 0 to 1, referring to how much that zone is used to start the possession or to end. If the weight of the zone is 1, the zone is highly related to the end of possessions. On the other side, if the weight is 0, the zone is related to the start of possessions. A color scale visually represents this weight.
- **Zone usage**. Each zone is then assigned a value depending on how much this zone and any arbitrary action is present in the heuristic network. Thus, this value refers to the importance of a certain zone to the purpose of the traces.
- **Zone dependencies**. The edges of the network are used to connect the zones and visualize the dependency between them.

5 Experiments

To demonstrate the ability of the proposed approach to infer and visualize team strategies for specific purposes in a game, we analyzed all team traces in which the attacking team successfully introduced the ball into the opponent's penalty box, and the possession did not originate from the last offensive third. After filtering and aggregating the data at the seasonal level, we present the most significant findings in the following sections. Firstly, we address the identification and representation of team-specific strategies; specifically, we investigate how teams penetrate their opponent's penalty box. Secondly, we evaluate the coherence of these strategies as an indicator of their resilience in executing their strategies. We assess the regularity with which teams adhere to these strategies throughout the season.

5.1 Data

The event data set was obtained from the work of [42], version 5. This dataset consists of a publicly available event data collection covering the 2017/2018 season for the Football first divisions of England, Spain, Germany, Italy, and France. The dataset includes 1,941 matches and more than 3 million events. The following examples utilized matches occurring in the English first division, including 380 matches. Concrete definitions and validation of the dataset can be accessed in the work of [42] and the supplementary online resource [43].

5.2 Discovering team strategies

The resulting event traces contain all the different ways each team could penetrate the opponent's box. Thus the discovery process aimed to identify patterns that could explain how each team approaches this task. This process yields two important artifacts, a Petri Net and a Heuristic Net. These logical artifacts can represent complex systems with high noise levels. However, such models can be difficult to interpret due to their complexity. Nonetheless, they are interesting artifacts for closer examination. In order to facilitate such examination, we have included examples of the generated models in the supplementary material. Here, we focus specifically on interpreting the resulting Heuristic maps and evaluation metrics.

The Petri Net contains the team's behavior toward the specified task. The Petri Net construct is useful to derive

TABLE 1
Fitness and generalization of the strategy models extracted from the top 10 teams in the English Premier League, 2017/2018 season.

League rank	Team	Fitness	Generalization
1	Manchester City	0.782	0.527
2	Manchester United	0.771	0.540
3	Tottenham Hotspur	0.732	0.547
4	Liverpool	0.770	0.527
5	Chelsea	0.735	0.521
6	Arsenal	0.785	0.536
7	Burnley FC	0.675	0.469
8	Everton	0.749	0.501
9	Leicester City FC	0.710	0.490
10	Newcastle United FC	0.729	0.485

conformance checking measures by utilizing the team traces against the discovered Petri Net. Table 1 presents the fitness and generalization metrics for the top 10 teams in the English Premier League during the 2017/2018 season. The table shows Manchester City had the highest fitness score (0.782), and Burnley FC had the lowest (0.675). Regarding generalization, Tottenham Hotspur had the highest score (0.547), and Burnley FC had the lowest (0.469). These results suggest that Manchester City had the most consistent and predictable behavior on the field, while Burnley FC had the most variable and difficult to predict behavior. Additionally, Tottenham Hotspur had a league-relative strong performance on fitness and generalization metrics, indicating a good balance between fitting to observed data and generalizing to new situations.

Furthermore, the Heuristic Net contains the flow of tasks and decisions in the team unfolding toward the opponents' penalty box. Thereafter, we can translate this information into Heuristic maps. Fig. 4 shows the process models extracted from Manchester City (MC) and Manchester United (MU). The causal relationships identify reasonable patterns if we analyze the Heuristic maps in detail. Manchester City has larger connectivity between zones denoting a high usage rate of all possible field parts and interconnecting them by their midfield players, including the areas closer to their own goal. On the contrary, Manchester United barely uses the closest areas to their goalkeeper, and their advance to more attacking positions is fixed to the center of the field. When reaching the midfield, MU utilizes all the width of the field. However, MU crosses the midfield more times to the flanks and less often to the center areas. This behavior is even clearer when advancing in the offensive third, where they use the flanks with higher frequency, especially the left attacking flank. MC shows different patterns in these zones; first, the midfield is approached mostly from the middle channels, and it is crossed quite equivalently to all channels with a small increase in the left side of the field. Second, the interconnections in the field's offensive half are more frequent and have larger possibilities.

5.3 Team strategy regularity

The discovery of these process models at the team level is difficult to evaluate as we need ground truth data. As this information is unknown, the efficiency of these models remains attached to how to interpret and use them. In this use case, we present an analysis to determine how regular teams are in their way of performing certain objectives (e.g., penetrating the opponents' penalty box). Thus, we refer to team strategy regularity as the ability of a team to remain resilient in some behavioral patterns throughout a set of games. We analyzed the seasonal models for the top 10 teams of the English Premier League in the 2017/2018 season and reviewed each round of the competitions. All the team purpose traces were replayed in the seasonal model for each round. After that, a round fitness value was obtained as the average fitness of all traces available in that round. A fitness of a trace is a value between 0 and 1 that refers to the model's ability to replay that trace. Fig. 5 shows the fitness by round and the overall fitness of each team. We also show ε as the mean square error between the fitness at every round and the overall season fitness. Interestingly, teams such as Manchester City or Arsenal show a high consistency of fitness in their matches which could lead to identifying these teams as resilient to their style of play or strategy. Conversely, teams like Manchester United or Tottenham show larger spikes in their fitness against the seasonal models denoting lower consistency in how they approach the task.

6 Discussion

The study of team sports is an area of increasing interest due to its potential for generating valuable insights into team strategy, match analysis, and player development. However, the analysis of large-scale event data in this field presents significant challenges, such as data variability and the complexity of team interactions. Process mining and visual analytics are emerging and promising approaches for addressing these challenges and unlocking valuable insights from sports big data sources. We propose a methodology that employs purpose-driven filters and field partitioning techniques to reduce the variance in football event sequences. These techniques enable us to focus on specific aspects of the game, such as attacking or defending, and to analyze the distribution of events on the field while keeping the logic underneath the team tactics. Using process discovery techniques, we extract logical artifacts that represent the team behavior in the field. These logical artifacts are then translated into Heuristic maps, a football-based visualization that allows for a detailed description of teams' event distribution on the field and dependencies between actions towards a certain objective.

The accuracy of the models generated by the methodology is evaluated in terms of fitness and generalization. Fitness refers to the ability of the model to accurately represent the observed data, while generalization refers to its ability to make accurate predictions about new data. While a more robust evaluation is impossible due to a lack of ground truth in the football domain and team tactics, the methodology and the discovered models' potential is demonstrated experimentally. The results show the potential of this approach for in-depth analysis of team behavior and how their strategy is implemented towards penetrating the opponent's box. In addition to providing insights into team strategy, this methodology can also be used to measure the regularity or resilience of a team to preserve a certain strategy of play

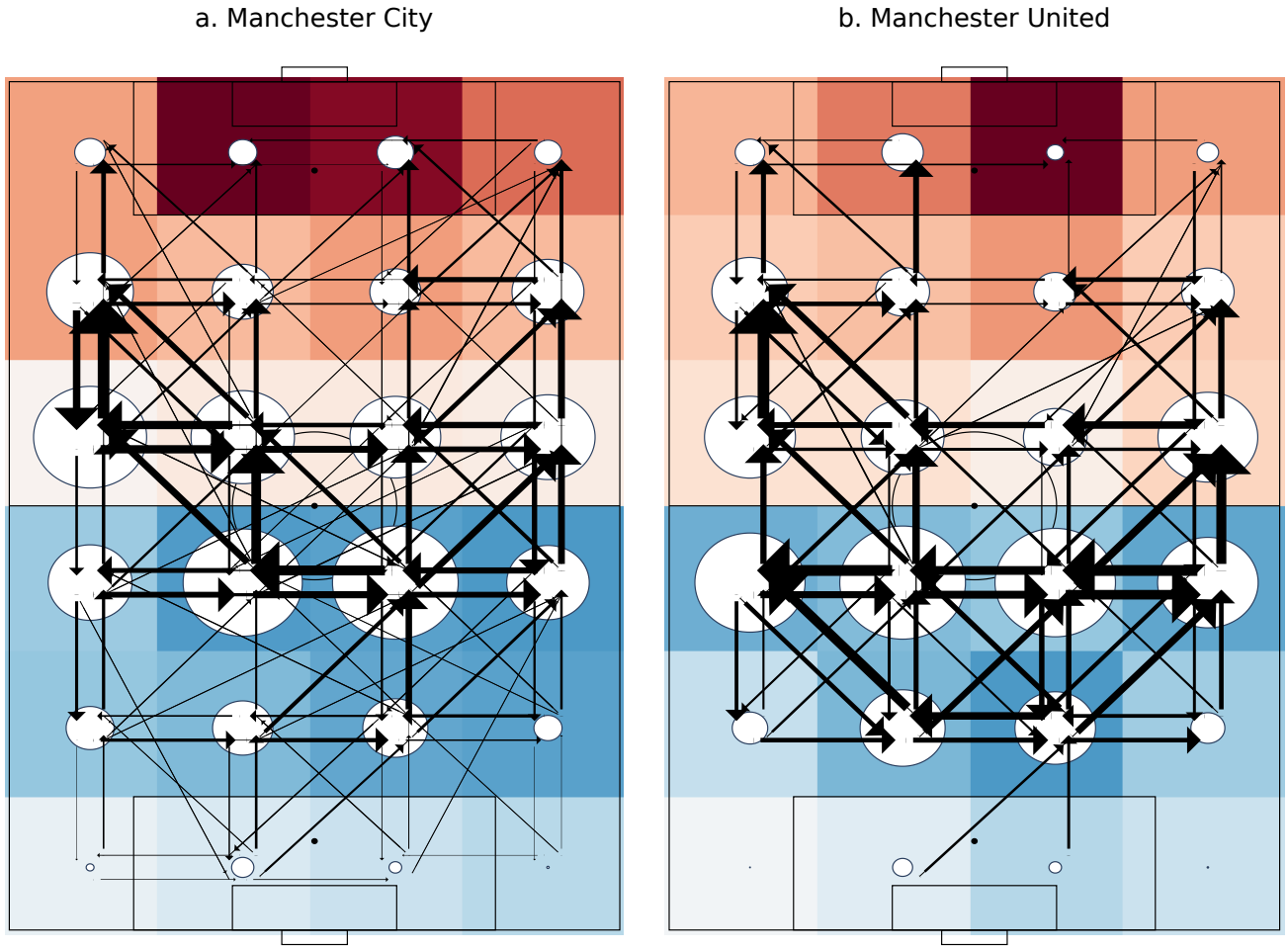

Fig. 4. Heuristic maps for Manchester City and Manchester United teams in the 2017/2018 season. First and second positions in the league, respectively. Relevant differences in both teams' tactics to penetrate the opponent's box are visible in terms of goalkeeper usage, midfield tendencies, and interconnections in the offensive third.

over the season. We can gain a better understanding of how teams adapt and evolve and how different strategies may be effective against opponents. This methodology could be combined with domain-specific analysis where different game states are considered and compared (e.g., differences depending on the match score or match context). While the focus of this analysis may not be directly related to success or finding productive patterns of play, the insights gained from the methodology can have significant implications for a team's ability to achieve its goals on the field. By understanding how teams execute their strategies and adapt to changing circumstances, coaches and analysts can develop more effective game plans and improve game execution, ultimately leading to greater success on the field. Most importantly, practitioners could develop predesigned process models representing the team plan or desired behavior and compare these models to the data-driven discovered models.

Other contributions have also investigated the use of event data to extract meaningful patterns and automatically document team tactics. These approaches usually focus on documenting tactics in success cases. Additionally, the sequentiality of the events is used to compute the probabilistic metrics and quantify team or player value. However, it is not proposed to explain the behavioral process of how the team unfolds during the game. [44] introduce an inductive logic approach to automatically discover patterns in successful offensive strategies without considering the event positions. In recent work, [34] clustered football sequences with similar spatiotemporal attributes and analyzed the shot-leading frequent patterns emerging while defining possession-based subsequences. In our approach, we extend these possession sequences with purpose-driven traces to reduce the variance in the sequences while being able to analyze the logical splits and team tactics. We also extend the current state of research by providing interpretable outputs in the form of logical artifacts such as Petri nets, Heuristic nets, and visual artifacts, introducing team Heuristic maps. Heuristic maps offer a complete view of team strategy for a given task, and they can be employed to analyze opponents' strategies or to validate team execution plans. This new visualization provides player interconnections and frequency of actions like passing networks. However, it increases its interpretability in motion-based team tactics by providing information about the starting and ending zones of the field and the dependencies and usages of each zone. Additionally, while ground truth validation is not possible, we provide evaluation metrics to measure the accuracy of the

identified tactics.

Some limitations also restrict the presented methodology. First and most importantly, the proposed methodology could be highly improved by integrating tracking data into the sequences. For instance, off-ball events could be automatically added to the event log to understand better each trace's logic and the overall process model. Also, the events lack contextual information about the other players' locations, which could lead to a better analysis. Regarding the usage of solely event data sequences, this paper could be extended by adding time-aware semantics to the process discovery. For instance, highlighting the time needed to perform a set of actions or identifying frequent paths shorter than others (i.e., fast-tempo moments of a football game). Furthermore, due to the inherent unsupervised nature of the pattern discovery task, the patterns are not subject to a ground truth validation. To further develop this methodology, it would be convenient to validate the findings with experts (i.e., coaches) and assess their value. Overall, this study provides a foundation for future research and innovation in the field of team sports analysis. Combining process mining and visual analytics techniques with domain-specific knowledge and expertise can unlock valuable insights from large-scale event data collections and gain a deeper understanding of player behavior and strategy in team sports.

SUPPLEMENTARY INFORMATION

The event logs used for Section 5 and examples of logical artifacts are attached to this manuscript in the form of supplementary material.

REFERENCES

[1] J. Gudmundsson and M. Horton, "Spatio-temporal analysis of team sports," *ACM Computing Surveys*, vol. 50, no. 2, pp. 1–34, apr 2017.

[2] H. Lepschy, H. Wäsche, and A. Woll, "Success factors in football: an analysis of the german bundesliga," *International Journal of Performance Analysis in Sport*, vol. 20, no. 2, pp. 150–164, feb 2020.

[3] R. Rein and D. Memmert, "Big data and tactical analysis in elite soccer: future challenges and opportunities for sports science," *SpringerPlus*, vol. 5, no. 1, aug 2016.

[4] W. Spearman, A. Basye, G. Dick, R. Hotovy, and P. Pop, "Physics-based modeling of pass probabilities in soccer," in *Proceeding of the 11th MIT Sloan Sports Analytics Conference*, 2017.

[5] G. Anzer and P. Bauer, "A goal scoring probability model for shots based on synchronized positional and event data in football (soccer)," *Frontiers in Sports and Active Living*, vol. 3, mar 2021.

[6] J. Davis, L. Bransen, L. Devos, W. Meert, P. Robberechts, J. Van Haaren, and M. Van Roy, "Evaluating sports analytics models: Challenges, approaches, and lessons learned," Jul. 2022.

[7] A. Hewitt, G. Greenham, and K. Norton, "Game style in soccer: what is it and can we quantify it?" *International Journal of Performance Analysis in Sport*, vol. 16, no. 1, pp. 355–372, apr 2016.

[8] T. Decroos, L. Bransen, J. V. Haaren, and J. Davis, "VAEP: An objective approach to valuing on-the-ball actions in soccer (extended abstract)," in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization, jul 2020.

[9] J. Fernandez-Navarro, L. Fradua, A. Zubillaga, P. R. Ford, and A. P. McRobert, "Attacking and defensive styles of play in soccer: analysis of spanish and english elite teams," *Journal of Sports Sciences*, vol. 34, no. 24, pp. 2195–2204, apr 2016.

[10] C. Diamantini, L. Genga, and D. Potena, "Behavioral process mining for unstructured processes," *Journal of Intelligent Information Systems*, vol. 47, no. 1, pp. 5–32, feb 2016.

[11] W. van der Aalst, A. Adriansyah, A. K. A. de Medeiros, F. Arcieri, T. Baier, T. Blickle, J. C. Bose, P. van den Brand, R. Brandtjen, J. Buijs, A. Burattin, J. Carmona, M. Castellanos, J. Claes, J. Cook, N. Costantini, F. Curbera, E. Damiani, M. de Leoni, P. Delias, B. F. van Dongen, M. Dumas, S. Dustdar, D. Fahland, D. R. Ferreira, W. Gaaloul, F. van Geffen, S. Goel, C. Günther, A. Guzzo, P. Harmon, A. ter Hofstede, J. Hoogland, J. E. Ingvaldsen, K. Kato, R. Kuhn, A. Kumar, M. L. Rosa, F. Maggi, D. Malerba, R. S. Mans, A. Manuel, M. McCreesh, P. Mello, J. Mendling, M. Montali, H. R. Motahari-Nezhad, M. zur Muehlen, J. Munoz-Gama, L. Pontieri, J. Ribeiro, A. Rozinat, H. S. Pérez, R. S. Pérez, M. Sepúlveda, J. Sinur, P. Soffer, M. Song, A. Sperduti, G. Stilo, C. Stoel, K. Swenson, M. Talamo, W. Tan, C. Turner, J. Vanthienen, G. Varvaressos, E. Verbeek, M. Verdonk, R. Vigo, J. Wang, B. Weber, M. Weidlich, T. Weijters, L. Wen, M. Westergaard, and M. Wynn, "Process mining manifesto," in *Business Process Management Workshops*. Springer Berlin Heidelberg, 2012, pp. 169–194.

[12] W. van der Aalst, *Process Mining*. Springer Berlin Heidelberg, 2016.

[13] G. Bergami, F. M. Maggi, A. Marrella, and M. Montali, "Aligning data-aware declarative process models and event logs," in *Lecture Notes in Computer Science*. Springer International Publishing, 2021, pp. 235–251.

[14] A. Kulakli and S. Birgun, "Process mining research in management science and engineering fields: The period of 2010–2019," in *Digital Conversion on the Way to Industry 4.0: Selected Papers from ISPR2020, September 24-26, 2020 Online-Turkey*. Springer, 2021, pp. 413–425.

[15] H. Mannila, H. Toivonen, and A. Inkeri Verkamo, "Discovery of frequent episodes in event sequences," *Data mining and knowledge discovery*, vol. 1, pp. 259–289, Sep. 1997.

[16] A. Stefanini, D. Aloini, E. Benevento, R. Dulmin, and V. Mininno, "A process mining methodology for modeling unstructured processes," *Knowledge and Process Management*, vol. 27, no. 4, pp. 294–310, jul 2020.

[17] A. Weijters, W. M. van Der Aalst, and A. A. De Medeiros, "Process mining with the heuristics miner-algorithm," *Technische Universiteit Eindhoven, Tech. Rep. WP*, vol. 166, no. July 2017, pp. 1–34, 2006.

[18] C. W. Günther and W. M. P. van der Aalst, "Fuzzy mining – adaptive process simplification based on multi-perspective metrics," in *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2007, pp. 328–343.

[19] S. J. J. Leemans, D. Fahland, and W. M. P. van der Aalst, "Discovering block-structured process models from incomplete event logs," in *Application and Theory of Petri Nets and Concurrency*. Springer International Publishing, 2014, pp. 91–110.

[20] M. Leemans and W. M. P. van der Aalst, "Discovery of frequent episodes in event logs," in *Lecture Notes in Business Information Processing*. Springer International Publishing, 2015, pp. 1–31.

[21] K. Guizani and S. A. Ghannouchi, "An approach for selecting a business process modeling language that best meets the requirements of a modeler," *Procedia Computer Science*, vol. 181, pp. 843–851, 2021.

[22] R. Dijkman, J. Hofstetter, and J. Koehler, *Business Process Model and Notation*. Springer, 2011, vol. 89.

[23] E. Sivaraman and M. Kamath, "On the use of petri nets for business process modeling," in *IIE Annual Conference. Proceedings*. Citeseer, 2002, p. 1.

[24] W. M. Van Der Aalst and A. H. Ter Hofstede, "Yawl: yet another workflow language," *Information systems*, vol. 30, no. 4, pp. 245–275, 2005.

[25] M. Garnica-Caparrós, "KPIs on the basis of match events data," in *Match Analysis*. Routledge, nov 2021, pp. 159–167.

[26] A. Bialkowski, P. Lucey, P. Carr, Y. Yue, S. Sridharan, and I. Matthews, "Large-scale analysis of soccer matches using spatiotemporal tracking data," in *2014 IEEE International Conference on Data Mining*. IEEE, dec 2014.

[27] ——, "Identifying team style in soccer using formations learned from spatiotemporal tracking data," in *2014 IEEE International Conference on Data Mining Workshop*. IEEE, dec 2014.

[28] P. Bauer and G. Anzer, "Data-driven detection of counterpressing in professional football," *Data Mining and Knowledge Discovery*, vol. 35, no. 5, pp. 2009–2049, jul 2021.

[29] B. Low, D. Coutinho, B. Gonçalves, R. Rein, D. Memmert, and J. Sampaio, "A systematic review of collective tactical behaviours in football using positional data," *Sports Medicine*, vol. 50, no. 2, pp. 343–385, sep 2019.

[30] B. Low, R. Rein, D. Raabe, S. Schwab, and D. Memmert, "The porous high-press? an experimental approach investigating tactical behaviours from two pressing strategies in football," *Journal of Sports Sciences*, vol. 39, no. 19, pp. 2199–2210, may 2021.
[31] A. Tenga and Ø. Larsen, "Testing the validity of match analysis to describe playing styles in football," *International Journal of Performance Analysis in Sport*, vol. 3, no. 2, pp. 90–102, dec 2003.
[32] J. Bekkers and S. Dabadghao, "Flow motifs in soccer: What can passing behavior tell us?" *Journal of Sports Analytics*, vol. 5, no. 4, pp. 299–311, dec 2019.
[33] L. Gyarmati, H. Kwak, and P. Rodriguez, "Searching for a unique style in soccer," *arXiv preprint arXiv:1409.0308*, 2014.
[34] T. Decroos, J. V. Haaren, and J. Davis, "Automatic discovery of tactics in spatio-temporal soccer match data," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, jul 2018.
[35] P. Kröckel and F. Bodendorf, "Process mining of football event data: A novel approach for tactical insights into the game," *Frontiers in Artificial Intelligence*, vol. 3, jul 2020.
[36] J. Clijmans, M. Van Roy, and J. Davis, "Looking beyond the past: Analyzing the intrinsic playing style of soccer teams," in *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases ECML PKDD 2022*, 2022.
[37] C. McCarthy, P. Tampakis, M. Chiarandini, M. B. Randers, S. Jänicke, and A. Zimek, "Analyzing passing sequences for the prediction of goal-scoring opportunities," in *Communications in Computer and Information Science*. Springer Nature Switzerland, 2023, pp. 27–40.
[38] N. S. N. Ayutaya, P. Palungsuntikul, and W. Premchaiswadi, "Heuristic mining: Adaptive process simplification in education," in *2012 Tenth International Conference on ICT and Knowledge Engineering*. IEEE, nov 2012.
[39] A. Berti, S. J. van Zelst, and W. van der Aalst, "Process mining for python (pm4py): Bridging the gap between process- and data science," 2019.
[40] J. C. A. M. Buijs, B. F. van Dongen, and W. M. P. van der Aalst, "Quality dimensions in process discovery: The importance of fitness, precision, generalization and simplicity," *International Journal of Cooperative Information Systems*, vol. 23, no. 01, p. 1440001, mar 2014.
[41] A. F. Syring, N. Tax, and W. M. P. van der Aalst, "Evaluating conformance measures in process mining using conformance propositions," in *Transactions on Petri Nets and Other Models of Concurrency XIV*. Springer Berlin Heidelberg, 2019, pp. 192–221.
[42] L. Pappalardo, P. Cintia, A. Rossi, E. Massucco, P. Ferragina, D. Pedreschi, and F. Giannotti, "A public data set of spatio-temporal match events in soccer competitions," *Scientific Data*, vol. 6, no. 1, oct 2019.
[43] L. Pappalardo and E. Massucco, "Soccer match event dataset," 2020.
[44] J. V. Haaren, V. Dzyuba, S. Hannosset, and J. Davis, "Automatically discovering offensive patterns in soccer match data," in *Advances in Intelligent Data Analysis XIV*. Springer International Publishing, 2015, pp. 286–297.

7 Biography Section

Marc Garnica-Caparrós is a Research Associate and PhD student at the German Sports University Cologne (Germany). His main expertise includes the design and modeling of software and database systems for data analytics. His main research interests are Big Data management and analysis systems and knowledge discovery applications in sport performance. His doctoral project leverages conceptual modeling and data analytics to improve the efficiency and interpretability of sports analytics scenarios. He received his BSc in Computer Science from the Universitat Politécnica de Catalunya, Spain and the Joint Master Degree in Big Data Management and Analytics.

Dilyar Iskan is a Research Assistant in the German Sports University of Cologne and a master student at RWTH Aachen. His main research interests involve Process Mining and Machine Learning models to create solutions for real-world challenging problems. He received his BSc and MSc in Computer Science from the RWTH Aachen, Germany.

Daniel Memmert is a Professor and Executive Head of the Institute of Exercise Training and Sport Informatics at the German Sport University Cologne, Cologne (Germany), with a visiting assistant professorship 2014 at the University of Vienna (Austria). Memmert received his PhD (basic cognition in team sports) and habilitation (creativity in team sports) in sport science from the Elite University of Heidelberg. In 2010 he was awarded 3rd place with Germany's most renowned German Olympic Sports Confederation (DOSB) Science Award. His research is focused on computer science in sports (pattern identification and simulation), human movement science (cognition and motor activity), sport psychology (attention and motivation), talent, children and elite research and research methods.

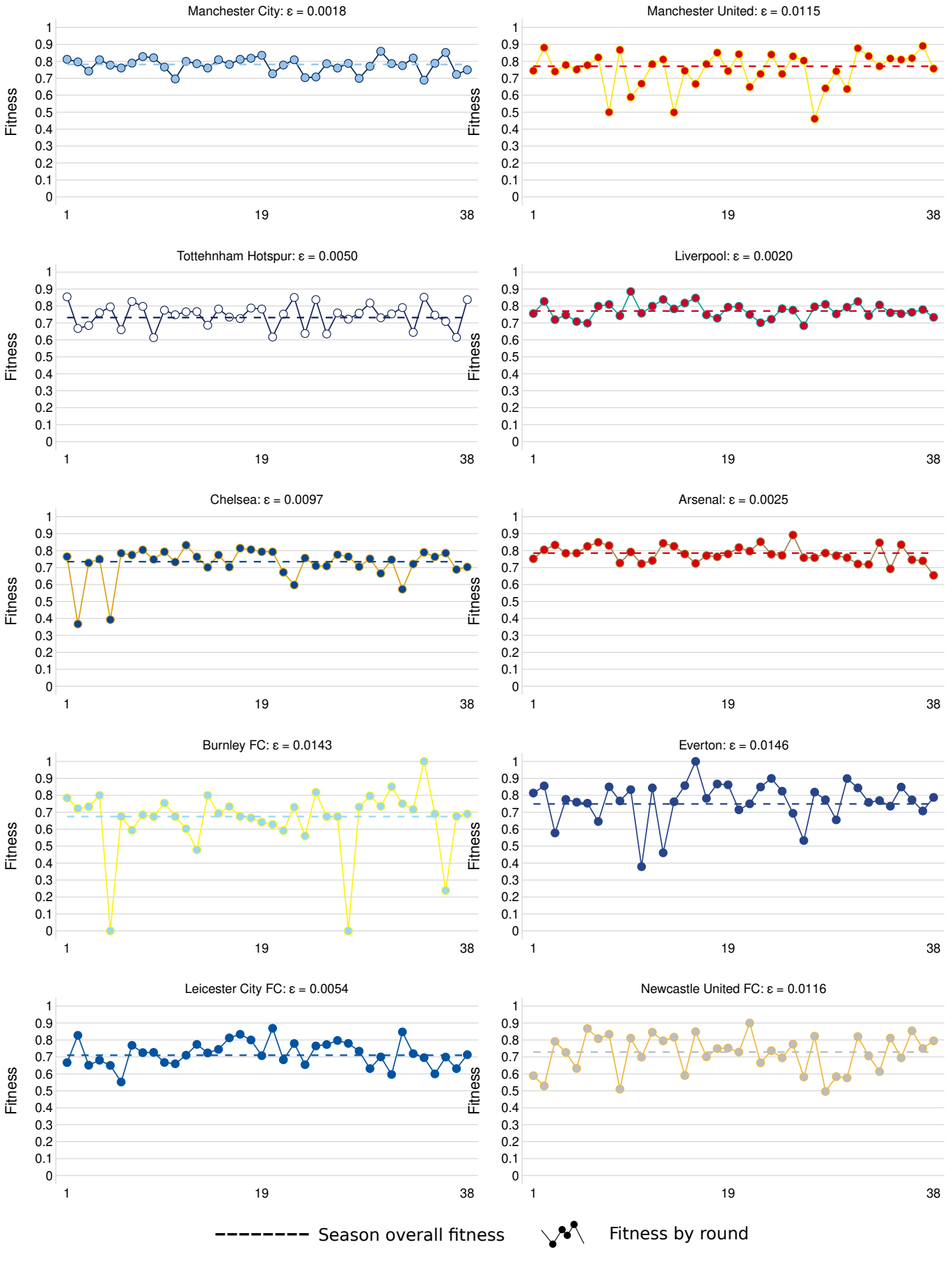

Fig. 5. Team strategy regularity of the top 10 teams of the English Premier League in 2017/2018. Season team strategy is evaluated at every round. Round fitness indicates whether each round's team traces align with the identified strategy at the end of the season. ε is the mean square error between the fitness at every round and the overall fitness. Examples of regular teams are Manchester City, Liverpool or Arsenal. Conversely, teams such as Manchester United, Tottenham, or Chelsea show more irregularity in their strategies.