Leveraging team traces to optimize process-aware tactical style discovery

Marc Garnica Caparrós
Institute of Exercise Training and Sport Informatics, German Sport University Cologne, Germany

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 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 reporting at high granularity all on-ball actions occurring during a game. 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]. The ideal scenario would consist of synchronizing the data streams to gain the best of both data contexts.

Hewitt et al. (2016) [4] define team playing style as “the characteristic playing pattern demonstrated by a team during games”. This inherent yet unknown aspect of football teams is essential in understanding the analytical outcomes. For instance, when assessing a potential new player for a team, it is desirable to know the 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.

Despite the media presenting football style 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 [5]. Additionally, an

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

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 use of advanced data sources in the realm of football analytics has mainly focused on the evaluation of actions and players. Nowadays, several methods allow for an objective measure of the impact of each action performed in the field, for instance, with probabilistic classification approaches toward reaching a certain reward (i.e., scoring opportunity with the widely presented Expected Goals metric). 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 the footprint or trace of such a system. Therefore, the presented methodology uses the possessions encoded in event data as traces of the team strategy. These event traces are then analyzed to identify common structures in which actions are taken towards a certain goal.

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 [6] 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 lack of structure in the traces, highlight frequent patterns of play, 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 and performance during a game. The methodology is evaluated in the 2021/2022 season of the English Premier League.

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.

2.1 Process mining unstructured systems

Process Mining (PM) is a semiautomatic evidence-based methodology to discover, monitor, evaluate, and improve processes [7]. 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 [8]. These process models might come in several forms. 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) [9]. These methods leverage the variants and deviations that might occur between the occurring events and the process model underneath [10].

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 [11], 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.

However, PM popularity in industry and research has increased in recent years from various applications and domains [10], processes no longer must have a predefined structure, and they combine elements from digital information systems with events captured from real environments. Therefore, the model discovering algorithm cannot assume 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. Stefanini et al. (2020) [12] 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-frequency behavior. The Heuristic Miner [13] tries to deal with the noise in the data and the high variety between logs by applying some heuristics to infer the underlying process model. Similarly, Günther et al. (2007) [14] 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 Leemans et al. (2014) [15] 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 Leemans et al. (2015) [16] for a comparison and detailed explanation of the existing algorithms for sequence mining and process-aware discovery.

The visual representation of the discovered models also plays an essential role in the discovery training and evaluation. A process discovery model's output is represented visually and conceptually by means of a process modeling language [17]. These languages can be notation-based graphical representations of workflows like BPMN [18] or mathematical models representing discrete variables, states, and transitions such as Petri Nets [19].

2.2 Team strategy data-driven identification in football

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]. Hewit et al. (2016) [4] 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 [20]. Football team formations describe the roles of each player on the field and are highly coupled with team tactics and strategy [21]. Team tactics are expected to include guidelines and rules for every game phase. 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 [22]. 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 [23]. Tracking data also can help shed light on the collective behavior of teams depending on their defensive strategies [24]. 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.

Most advanced approaches acknowledge the sequentiality of the event data source. On the one hand, methods like VAEP [25] 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's 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

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

involved, and location on the field. Bekkers et al. (2019) [26] extend the concept of network motifs applied to passing sequences in football [27] to provide a tactical and statistical analysis of passing behavior in football. However, these motifs do not include any spatiotemporal information about the events, and the time between passes is restricted. However, in a process-aware analysis, the time between activities is taken into account for the process discovery [8]. In a similar approach to this paper [28], the authors apply clustering techniques and a success score metric to discover frequent sequential patterns at the team level. Sequential patterns are also studied in buildup plays [29]. The authors use a grid division as well as additional game states stemming from the availability of teammates and defenders to build a Markov decision process and visualize the most frequent combinations in buildup scenarios.

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]. Kröckel et al. (2020) [30] 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.

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 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 purpose-driven sequences

The proposed methodology for processing event data, identifying purpose-driven team sequences, and discovering team patterns is displayed in Figure 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

[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

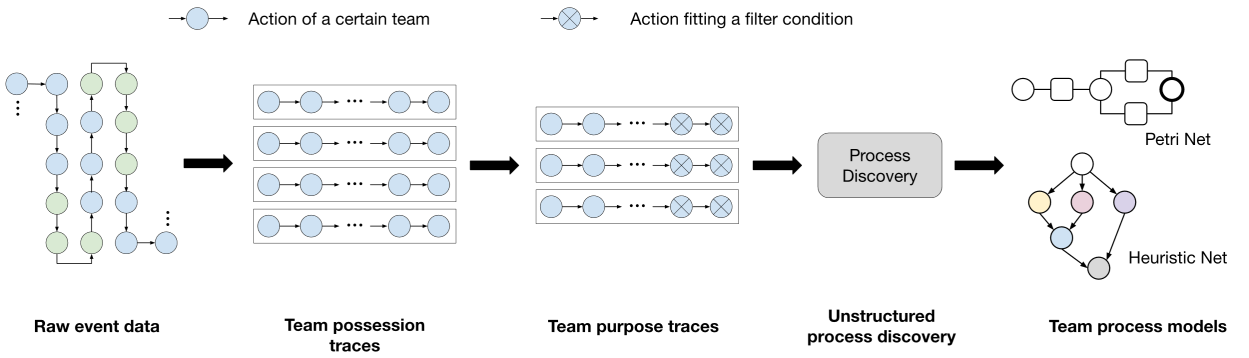

Figure 1: Process Discovery methodology. Raw event data is mapped into team trace logs containing the sequences performed by each team. Certain criteria must be modeled to filter these traces to remove noise and construct the team purpose traces with similar sequences (similarity defined by arbitrary 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.

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 possession of the ball.

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

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 [31, 32, 29]. An example of a simple field partitioning with highlighted common cells between two event

sequences is shown in Figure 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. Other data-driven split criteria are available, such as hierarchical partitioning or trace clustering [28].

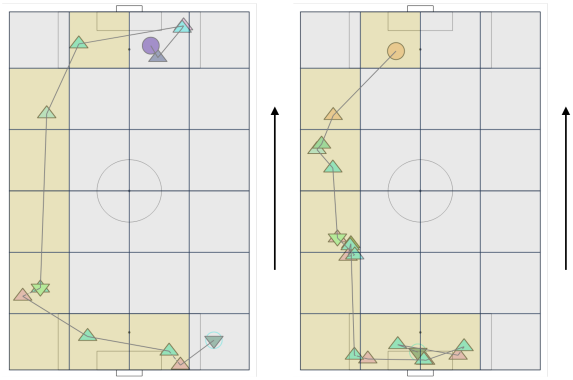

Figure 2: A simple field partitioning helps highlight the common zones the team used to build the sequence up and access the opponent penalty box.

Discovery of team processes

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 (HM) algorithm. For each trace, several attributes are described according to the conceptual model so that the discovery can be executed.

The HM algorithm is based on the construction of a dependency graph. The dependency graph is a frequency-based data structure that gathers a level of certainty of the dependencies between events present in the data. Once this structure is constructed, the process model is inferred from these dependencies, for instance, activities are identified that directly follow other events, and parallel flows are identified if two activities occur very often together and have similar dependencies. Finally, the algorithm also detects loops of activities and adds them to the model. The model can take various forms or syntaxis, for instance, a direct graph or a Petri Net. The resulting model can be refined by adjusting the relations between events or removing unnecessary activities. We refer to Weijters et al. [13] for a complete technical description of the HM algorithm.

Several attributes are mandatory for the algorithm. The case identifier 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 identifier 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 identifier. The activity identifier 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. We base our discovery approach on the Heuristic Miner (HM) [13, 33]. 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%. Thus, the algorithm considers dependency values over 50% to ensure we are not capturing low-frequency 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 [34].

4.1 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 the work of Buijs [35] and Syring [36] for a detailed description and comparison of these two metrics.

4.2 Heuristic Maps: Translating process models to football

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. This logical outcome can be explored directly from the result of the HM algorithm.

However, translating these models into practical information takes time and effort. Therefore, we also present 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, its content is based on the information represented for the causal network produced by the HM. A detailed

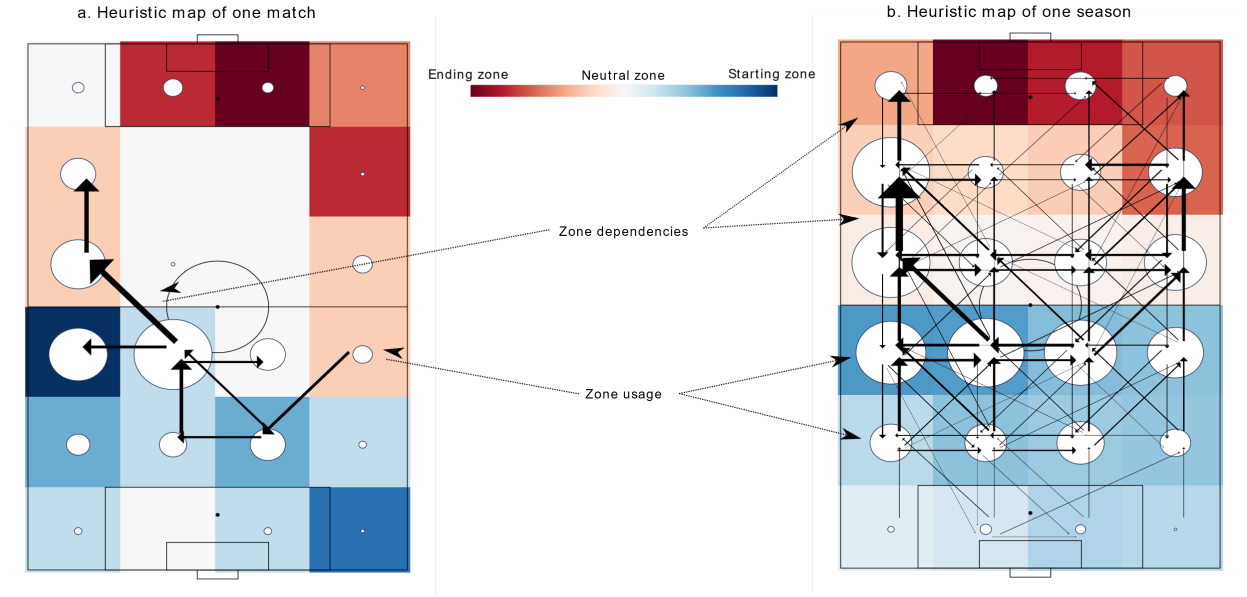

Figure 3: Two examples of Heuristic maps with their components, a Heuristic Map of one match (a) and a Heuristic map of the same team for a whole season (b).

explanation of the proposed Heuristic maps visualization is presented in Figure 3. 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.

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 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. However, such models can be difficult to interpret due to their complexity. Nonetheless, they are interesting artifacts for closer examination.

Table 1: Fitness and generalization of the team style models extracted from 6 teams of the English Premier League, 2021/2022 season

Team	Fitness	Generalization
Manchester City	0.877	0.703
Arsenal	0.818	0.712
Manchester United	0.814	0.701
Liverpool FC	0.828	0.717
Tottenham Hotspur	0.785	0.733
Chelsea	0.820	0.707

Table 1 presents the fitness and generalization metrics for the top six teams in the English Premier League during the 2021/2022 season. The table shows Manchester City had the highest fitness score (0.877), and Tottenham had the lowest (0.785). Regarding generalization, Tottenham Hotspur had the highest score (0.733), and Manchester United had the lowest (0.701). These results suggest that Manchester City had the most consistent and predictable behavior on the field, while Tottenham 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.

We can translate the logical models into Heuristic maps. Figure 4 shows the process models extracted from Manchester City (MC), Arsenal, Manchester United (MU), and Liverpool. The causal relationships identify reasonable patterns if we analyze the Heuristic maps in detail. MC 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, teams like MU or Liverpool barely use 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. Similarly, Liverpool also uses all the width of the field to cross the midfield. 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 and Arsenal show 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 for Arsenal. Overall, these maps aim to translate the logical findings by the process discovery algorithm. They give a general overview on how teams unfold their process towards the specific objective.

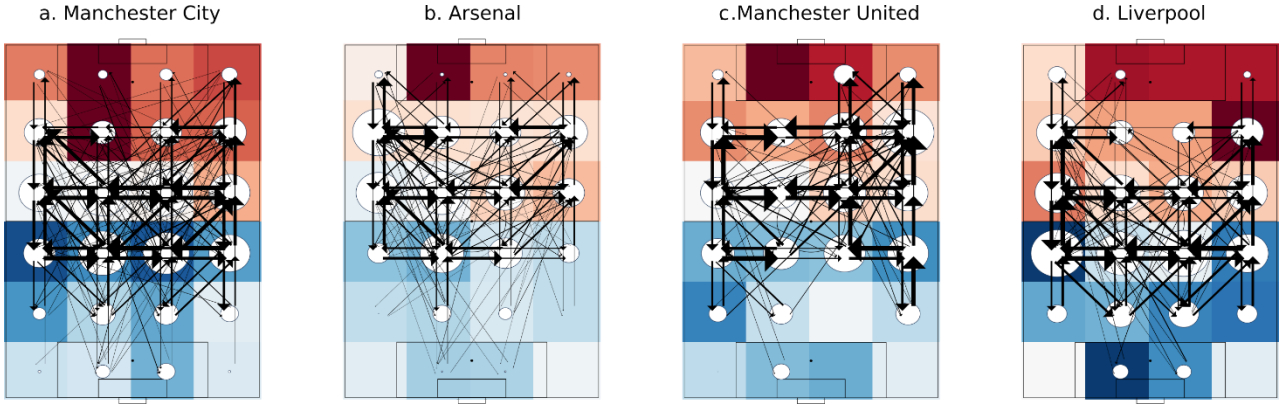

Figure 4: Seasonal Heuristic maps for English Premier League teams Manchester City (a), Arsenal (b), Manchester United (c) and Liverpool (d).

5.2 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 six teams of the English Premier League with data from the 2021/2022 season and reviewed each round of the competitions. All the team purpose traces were replayed in the seasonal model for each round. Therefore, the execution of an individual match was compared to the overall inference of the team strategy at the end of the season. The fitness of a trace is a value between 0 and 1 that refers to the model's ability to replay that trace. Determining whether a trace is replayable by a model refers to checking whether a given sequence of events (the trace) can be successfully and completely executed based on the rules and structure defined in the process model. This usually involves simulating the execution of the trace within the process model, ensuring that each step in the trace can be executed according to the model's rules. If the entire trace can be replayed without encountering conflicts or violations of the model's structure and constraints, it is considered replayable. Figure 5 shows the fitness by round and the overall fitness of each team. Matches with higher fitness mean that the actions unfolding on these matches are correctly represented by the seasonal models. Lower fitness values indicate infrequent behavior in those matches. We also computed the mean square error between the fitness at every

round and the overall season fitness (ε). 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.

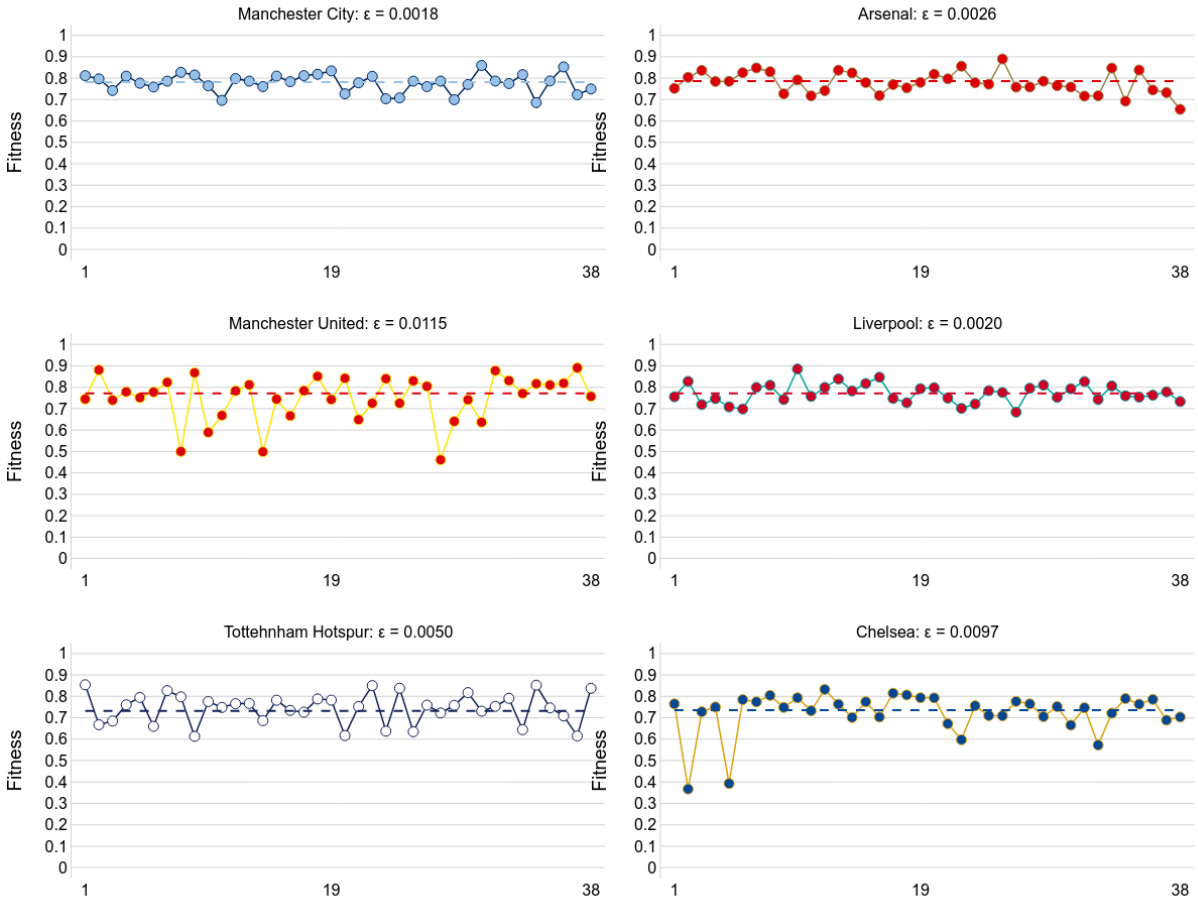

Figure 5: Team strategy regularity of the six teams of the English Premier League in 2021/2022. Season team strategy is evaluated at every round. Round fitness indicates whether each round's team traces align with the identified strategy at the end of the season. ε is the mean square error between the 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.

Discussion

The analysis of large-scale event data in this field presents significant challenges, such as data variability and the complexity of team interactions for sequential modeling and pattern mining. 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 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 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.

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 better understand 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 [29]. 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 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, 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 for different purposes.

Acknowledgments

This research was supported by the German Sport University of Cologne. Data was provided by StatsBomb.

References

[1] J. Gudmundsson and M. Horton, "Spatio-Temporal Analysis of Team Sports," *ACM Computing Surveys,* vol. 50, p. 1–34, April 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, p. 150–164, February 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, August 2016.

[4] 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, p. 355–372, April 2016.

[5] 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, p. 2195–2204, April 2016.

[6] C. Diamantini, L. Genga and D. Potena, "Behavioral process mining for unstructured processes," *Journal of Intelligent Information Systems,* vol. 47, p. 5–32, February 2016.

[7] 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, p. 169–194.

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

[9] 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, p. 235–251.

[10] 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*, 2021.

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

[12] 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, p. 294–310, July 2020.

[13] A. J. M. M. Weijters, W. M. P. van Der Aalst and A. A. De Medeiros, "Process mining with the heuristics miner-algorithm," *Technische Universiteit Eindhoven, Tech. Rep. WP,* vol. 166, p. 1–34, 2006.

[14] 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, p. 328–343.

[15] 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, p. 91–110.

[16] 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, p. 1–31.

[17] 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, p. 843–851, 2021.

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

[19] E. Sivaraman and M. Kamath, "On the use of Petri nets for business process modeling," in *IIE Annual Conference. Proceedings*, 2002.

[20] 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*, 2014.

[21] A. Bialkowski, P. Lucey, P. Carr, Y. Yue, S. Sridharan and I. Matthews, "Identifying Team Style in Soccer Using Formations Learned from Spatiotemporal Tracking Data," in *2014 IEEE International Conference on Data Mining Workshop*, 2014.

[22] P. Bauer and G. Anzer, "Data-driven detection of counterpressing in professional football," *Data Mining and Knowledge Discovery,* vol. 35, p. 2009–2049, July 2021.

[23] 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, p. 343–385, September 2019.

[24] 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, p. 2199–2210, May 2021.

[25] 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*, 2020.

[26] J. Bekkers and S. Dabadghao, "Flow motifs in soccer: What can passing behavior tell us?," *Journal of Sports Analytics,* vol. 5, p. 299–311, December 2019.

[27] L. Gyarmati, H. Kwak and P. Rodriguez, "Searching for a unique style in soccer," *arXiv preprint arXiv:1409.0308,* 2014.

[28] 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*, 2018.

[29] M. Van Roy, P. Robberechts and J. Davis, "Optimally Disrupting Opponent Build-ups," in *StatsBomb Conference*, 2021.

[30] 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, July 2020.

[31] 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.

[32] 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, p. 27–40.

[33] 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*, 2012.

[34] A. Berti, S. J. van Zelst and W. van der Aalst, *Process Mining for Python (PM4Py): Bridging the Gap Between Process- and Data Science,* arXiv, 2019.

[35] 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, p. 1440001, March 2014.

[36] 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, p. 192–221.