Looking Beyond the Past: Analyzing the Intrinsic Playing Style of Soccer Teams

Jeroen Clijmans, Maaike Van Roy (✉), and Jesse Davis

Department of Computer Science & Leuven.AI, KU Leuven, Belgium
jeroen.clijmans@student.kuleuven.be
maaike.vanroy@kuleuven.be
jesse.davis@kuleuven.be

Abstract. Analyzing the offensive playing style of teams is an important task within soccer analytics that has various applications in match preparation and scouting. Existing data-driven approaches typically quantify style by looking at individual events that occur during a match in isolation. This approach has two shortcomings. First, it ignores the sequential aspect of the game, as patterns of play are a crucial aspect of playing style. Second, it fails to generalize over the limited amount of data in order to model slight variations of the observed patterns that a team may employ in the future. This is particularly important when considering rare actions like shots and goals, which are the key success criteria of an offensive style. This paper proposes a novel approach for analyzing playing style that addresses these shortcomings. First, it captures the sequential patterns of a team's style by modeling the observed behavior of a team as a discrete-time Markov chain. Second, it characterizes the offensive style of teams in a number of features that are based on domain knowledge. It applies a combination of analytical techniques and probabilistic model checking to reason about a team's model in order to extract values for these features. As the model allows for a generalization of a team's past behavior, the extracted style is less influenced by the rarity of shots and goals. Using event stream data of the 2019/20 English Premier League, we empirically show that the proposed approach can capture a team's positional and sequential style, as well as reason about the style's efficiency and similarities with other teams.

Keywords: Markov Model · Probabilistic Model Checking · Playing Style · Soccer Analytics.

1 Introduction

Analyzing the in-game behavior of teams (i.e., their playing style) has several important use cases in professional soccer. For example, identifying a team's typical patterns of movement or strategies can be used to aid match preparations such as designing a game plan that exploits the weaknesses of the opponent, or scheduling pre-tournament friendlies based on the similarity between the pre-

tournament and in-tournament opponent's playing styles.[1] Additionally, it can also be used for player acquisition, where a club may be interested in targeting players that currently play for a team that is stylistically similar.

Consequently, an important question is how to characterize a team's style of play. One way to do this is based on manual video analysis of matches. However, this is inherently subjective and time-consuming, making it impossible to do this for a large number of matches or teams. Hence, a data-driven approach can play a role by, for example, identifying a shortlist of teams most similar to an upcoming opponent or identifying insights that are difficult for humans to pick up on. Existing approaches mainly focus on quantifying style at the level of individual events in a match [3,5,7]. This has two important limitations. First, it ignores the sequential nature of the game which is crucial for modeling patterns of play. Second, it fails to generalize over the limited amount of data in order to capture slight variations of the observed patterns that a team may employ in the future. As a season is relatively short and players rarely perform the exact same actions multiple times, the data is inherently limited. Using data of previous seasons is often not useful, as changes in players and management, and thus in style, happen regularly. Especially when analyzing the offensive style of teams, in which rare actions such as shots and goals play an important role, being able to generalize over the limited amount of data and capturing the *intrinsic* playing style of teams becomes particularly important.

This paper proposes a novel approach for playing style analysis based on a learned model that captures a team's intrinsic offensive behavior from historical event stream data.[2] In particular, we model the behavior of a team as a discrete-time Markov chain (DTMC). This has the inherent advantages that the sequential nature of the game is taken into account and observed patterns are interleaved allowing for generalization beyond past behavior. Additionally, we define a number of features that characterize playing style based on domain knowledge. Intuitively, these features capture how often teams employ certain stylistic parameters and how effective they are doing so. Then, we show how a combination of analytical techniques and probabilistic model checking can be used to reason about each team's learned model to obtain values for the features we defined, thereby characterizing their intrinsic playing style.

We illustrate our approach on event stream data of the 2019/20 English Premier League. Our approach indicates that Manchester City is the least likely to launch a counterattack and the most likely to eventually arrive at a shot by using combination play; Bournemouth should have considered using their left side more often, as the model considers it a side from which much more danger could have been created than the right side; and Leicester City's playing style was, out of all smaller teams, the most similar to the possession-based playing style that is often employed by big clubs such as Manchester City and Liverpool.

[1] `https://www.reuters.com/article/socccer-euro-bel/soccer-belgium-coach-martinez-outlines-euro-2020-warm-up-plans-idUKL8N29V0Q4`

[2] The implementation is publicly available: `https://github.com/JeroenClijmans/MarkovSoccer`

2 Capturing Team Behavior as a DTMC

The goal of this work is to capture and characterize the intrinsic playing styles of soccer teams. To this end, we propose to model the in-game offensive behavior of each team using a team-specific discrete-time Markov Chain (DTMC). Specifically, this model represents the behavior of the team during a possession sequence and will be learned from the team's historical on-the-ball actions. Next, we describe the data set used and outline the models and how they can be learned from historical data.

2.1 Data set

The models are constructed using historical event stream data. This type of data typically contains all on-the-ball actions (e.g., passes, dribbles, shots) that occur during a match and records various features about these actions such as location, involved players, timestamp, etc. In this work, we use event stream data from the 2019/20 English Premier League, which consists of 380 matches. We encode this data set to SPADL[3], which is a vendor-independent format to describe on-the-ball player actions and which facilitates the analysis [4].

2.2 Retrieving possession sequences

As a first step, before constructing the models, we extract all possession sequences from the data. We exclude possession sequences resulting from corners, crossed free-kicks, goal attempts from free-kicks and penalties, as these often involve custom tactics that are beyond the scope of this work.

We define a possession sequence as a maximal uninterrupted sequence of consecutive actions by the same team that either 1) starts with an action bringing the ball into play (e.g., a throw-in), or 2) involves three or more deliberate ball-moving actions[4] by the team under consideration. The former indicates that the team surely has control over the ball as it signifies the start of a possession sequence. The latter indicates that the team has *established* ball control during the sequence, otherwise they would be unable to execute these actions.

2.3 Constructing team-specific DTMCs

The extracted possession sequences of each team are used as input to learn each team's model. Specifically, the proposed model captures (1) how and where the ball is gained, (2) where the team tends to move the ball to, and (3) how and where the possession sequence eventually ends. The model is schematically sketched in Figure 1 and is defined by the following set of states and transitions:

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

[4] We define a deliberate ball-moving action as an action in which the main objective is to *deliberately move the ball to a certain position*. This includes actions such as passes, crosses, carries, and shots, but excludes actions such as clearances.

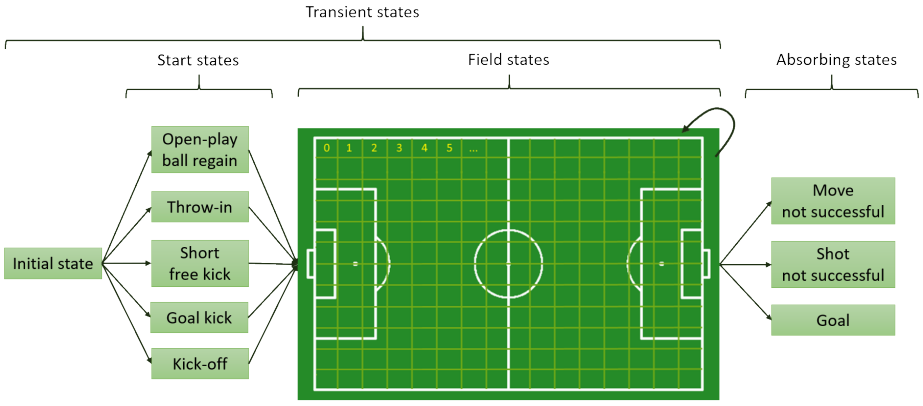


Fig. 1: Schematic overview of the states and transitions of the DTMC used to model the in-game offensive behavior of a team.

Transient states $\mathcal{T}$ This set of states can be entered and exited during a possession sequence. We define two types of transient states: *start states* and *field states*. *Start states* represent how possession of the ball is gained. We include five types: a throw-in, a short free-kick, a goal kick, a kick-off, and an open-play ball regain. *Field states* represent the particular locations on the pitch in which the ball can be situated during the possession sequence. We use discretized locations and divide the field up into 192 field states using a 12×16 grid.

Absorbing states $\mathcal{A}$ This set of states cannot be left once entered and indicates how a possession sequence can end. We include a *move not successful* (mns), a *shot not successful* (sns), and a *goal* (g) state, which represent losing the ball when trying to move it to another location, an unsuccessful shot and a successful shot, respectively.

Transitions How a team moves the ball from one state to another during a possession sequence is modelled by the transitions. Each transition is associated with a probability, which corresponds to the frequency of the corresponding action in the extracted sequences of the team. Concretely, we include the below transitions and calculate the probabilities as follows:

- **Initial state $init$ to start state s_t:** These probabilities are calculated as $P(s_t|init) = c_{seq,s_t}/c_{seq}$ where c_{seq,s_t} is the number of sequences starting with an action corresponding to the start state of type t and c_{seq} is the total number of sequences.
- **Start state s_t to field state f_i:** These probabilities are calculated as $P(f_i|s_t) = c_{s_t,f_i}/c_{s_t}$ where c_{s_t,f_i} is the number of actions that correspond to the start state of type t and after which the ball ends up in state f_i, and c_{s_t} is the total number of actions that correspond to the start state of type t.

- **Field state f_i to field state f_j:** These probabilities are calculated as $P(f_j|f_i) = c_{f_i,f_j}/c_{f_i}$ where c_{f_i,f_j} is the number of ball-moving actions starting in state f_i that successfully end up in state f_j, and c_{f_i} corresponds to the total number of actions (i.e., failed or successful) initiated in state f_i.
- **Field state f_i to an absorbing state:** Actions from field states can also result in the end of the sequence (i.e., failed actions or a goal). The probabilities for these transitions are calculated as $P(mns|f_i) = c_{f_i,mns}/c_{f_i}$, $P(sns|f_i) = c_{f_i,sns}/c_{f_i}$, and $P(g|f_i) = c_{f_i,g}/c_{f_i}$. Here, $c_{f_i,mns}$ is the number of unsuccessful ball-moving actions from state f_i, $c_{f_i,sns}$ is the number of unsuccessful shots from f_i, $c_{f_i,g}$ is the number of goals scored from f_i, and c_{f_i} is the total number of actions initiated in f_i.
- **Absorbing state:** This is equal to a self loop with probability one.

Using a DTMC confers several advantages for trying to capture a team's style of play. First, DTMCs go beyond considering a single action in isolation by modeling sequences of consecutive actions. As interplay between actions is important, this gives a more comprehensive perspective on a team's style. Second, they are able to generalize over the different actions that a team has performed in the past. Thus, these models allow us to reason about ways in which the team could combine different actions during a possession sequence, even though it was not explicitly observed in the data. This also means that the model is less influenced by the rarity of events such as shots or goals.

3 Characterizing a Team's Playing Style

From a soccer perspective, there are a number of potential behavioral patterns and characteristics of play that are relevant and indicative of style such as:

Preference for certain locations. Teams like to work the ball through certain zones on the pitch. This may arise due to tactical instructions, such as Manchester City's use of half spaces, or because teams have strong players in certain positions, which may manifest itself as preference for using one side of the pitch.

Preference for certain sequences. Teams like to use and reuse various combinations of actions which allow them to move the ball between locations. Some teams have a preference for playing the ball wide and down the flanks, while other teams predominantly use sequences going through the center of the field.

Directness of play. Some teams like Manchester City will employ a patient and structured style that methodologically tries to build up to a goal scoring opportunity. Other teams like Everton like to sit deeper and rely on a more direct counter attacking style of play.

Ability to create shots. Generating (high-quality) shots is extremely important and is often done by employing particular patterns of play. Capturing how

effective teams are at generating shots from various locations on the pitch can give some indications of style.

For each of the aforementioned categories we define a number of different features that can be computed by reasoning about our learned model of a team's behavior. Intuitively, each feature either captures how often a team employs a strategy or how effective a team is at applying a given strategy. These features can also capture relative strengths of a team such as their effectiveness of generating shots when attacking from the left vs. right flank. Next, we describe for each category the different features we defined and how they can be computed using the model.

3.1 Features regarding a team's preference for certain locations

One indicator of style of play is a team's preference for working the ball into certain locations. We consider a team's locational preferences in two situations: general possession sequences, and more promising possession sequences that end with a shot. These two situations provide insights into both a team's regular playing style and their style when playing in a more successful manner. Using these situations, we derive six features that allow us to characterize a team's playing style based on their preferred locations. We compute these features in two steps. First, we construct heatmaps containing the expected number of times a team will possess the ball in each location of the pitch during both situations. Second, using these heatmaps, we derive three concrete features for each situation indicating a team's preference for the left, right, and middle part of the pitch.

Step 1: Constructing heatmaps: We compute the expected number of visits to each location using the fundamental matrix N of the model:

$$N = \sum_{k=0}^{\infty} Q^k = (I - Q)^{-1}. \tag{1}$$

Here, I is the identity matrix and each entry q_{ij} of Q is the transition probability from transient state i to transient state j. Each entry n_{ij} is equal to the expected number of visits to transient state j when starting the possession sequence from state i. We compute this matrix both for general possession sequences as well as for only those ending with a shot. Computing the fundamental matrix $\hat{N}$ that solely generates those sequences of the original model which end in a shot requires a slight alteration to the model. More specifically, we restrict the set of absorbing states $\hat{\mathcal{A}}$ to only include the *shot not successful* and *goal* states. The fundamental matrix $\hat{N}$ of the new model can be computed as in Equation 1 with the new transition probabilities from transient to transient states given by the matrix $\hat{Q}$:

$$\hat{Q} = D_0^{-1} Q D_0. \tag{2}$$

Here, D_0 is the diagonal matrix with, for each transient state i, an entry $b_{i\hat{a}}$:

$$b_{i\hat{a}} = \sum_{j \in \hat{\mathcal{A}}} b_{ij} \tag{3}$$

$$B = NR \tag{4}$$

with N the fundamental matrix of the original model and R the matrix containing the original transition probabilities from transient to absorbing states.

The entries of the fundamental matrix yield a heatmap which is already interesting in its own right because it allows us to analyze which exact locations teams prefer to use. Additionally, contrasting a team's preference during general possession sequences with their preference during more promising ones will allow us to identify locations from which a team is more/less efficient.

Step 2: Computing features: Second, we derive three concrete features from each of the two heatmaps. More precisely, we derive a team's preference for the left side, the right side, and the central part of the field by calculating the relative percentage that the team uses each zone (illustrated in Figure 2), as given by the heatmaps. This gives an indication of which parts of the field a team tends to use more often (e.g., left side over the right side, or predominantly through the center), both during general play and when playing more successfully.

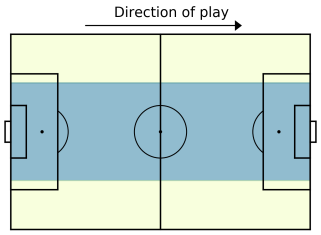


Fig. 2: Illustration of the three zones used to compute a team's locational preference. The left and right sides of the field (yellow) have a width equal to a quarter of the field. The remaining part of the field (blue) is defined as the central zone.

3.2 Features regarding a team's preference for certain sequences

A second indicator of style of play is a team's preference for combinations of consecutive locations, i.e., sequences. This provides insights into which locations of the pitch are often used together to move the ball from one location to the next. For example, whether a team prefers to move the ball wide and attack down the flanks or move it more centrally. We capture this aspect in two steps.

First, we generate the 200 most likely two-action sequences. The number of sequences to generate was empirically determined, and using more than 200 does not influence the results. The likelihood of a sequence is computed by multiplying the transition probabilities between the states of the sequence and weighing it by the expected number of visits to the first state of the sequence starting from the initial state. This weight is given by the fundamental matrix.

Second, we define two concrete features: *inward/outward preference*, which is the fraction of the 200 most likely sequences that move the ball inwards towards the middle of the pitch/outwards towards the touchlines.

3.3 Features regarding the directness of play

A third indicator of style is the directness of play. Namely, how fast a team tends to go from recovering the ball to creating a shooting opportunity, and how directly they do this. While some teams prefer a steady build up from the back using many short on-the-ball actions, other teams prefer to utilize long balls or to sit back and counter. We capture this aspect of style in four concrete features.

The first feature captures the team's speed of play during dangerous attacking sequences by measuring the average number of actions in a sequence that ends in a shot according to the model. We compute this feature in two steps. First, a new model is constructed which only takes into account possession sequences that end in a shot (see Section 3.1). Second, we use probabilistic model checking techniques (i.e., PRISM [10]) to compute the average number of actions in such sequences. The higher the number of actions needed, the more a team prefers a slower possession-based style over a faster direct style of play.

The second feature captures the team's probability of performing long goal kicks, which we define as goal kicks that end in the opponent half. This provides insights into a team's directness of build up play. We compute the probability of a team performing these long goal kicks by a summation over all transition probabilities that originate from the *goal kick* start state and end in any field state that is in the opponent's half.

The third feature captures the team's probability of performing long balls. We define these as actions that originate from the defensive half, bypass midfield, and end up in the final third of the pitch. The probability of a team performing these long balls can be computed by a weighted summation over all transition probabilities from states in the defensive half to states in the final third of the pitch. The weight assigned to each state is the relative usage of each location, given by the fundamental matrix. This is scaled so that the entries corresponding to the own half sum to 1, yielding a probability.

The fourth feature captures the team's probability of performing a successful counterattack. We define these as possession sequences that start in the team's own half, after an open-play ball regain, and yield a shot within eight actions. To calculate this probability, we first use a probabilistic model checker to compute the probability of arriving at a shot within eight actions for all locations in the own half. Next, we weight these locations by the probability of recovering the ball there, scaled so that these sum to 1.

3.4 Features regarding the ability to create shots

A final indicator of style that we consider is the team's ability to create shots. More specifically, we capture the team's probability of creating *non-opportunistic* shots. We define these as shots in a possession sequence that started in the team's own half. In contrast to generating a shot after recovering the ball in the opponent's half after they made a mistake, these shots better capture a team's ability to generate shots via smart ball movements. We capture this in two steps.

First, we compute a heatmap in which each entry is equal to the team's probability of generating a shot later on in the possession sequence when starting from the corresponding location on the field. This heatmap is computed using the formula $B = NR$, where N is the fundamental matrix and R the matrix containing all the transition probabilities from transient to absorbing states. As there are two absorbing states that entail that a shot has happened in our model (i.e., the *shot not successful* and *goal* state), the values of these two absorbing states are summed to calculate the final probability for each possible start state.

Second, we capture the non-opportunistic shot probability in one feature by means of a weighted average over all obtained probabilities for states in the team's own half. The weight of each location is the relative usage of that location, given by the fundamental matrix of the model and scaled so that the entries corresponding to the own half sum to 1.

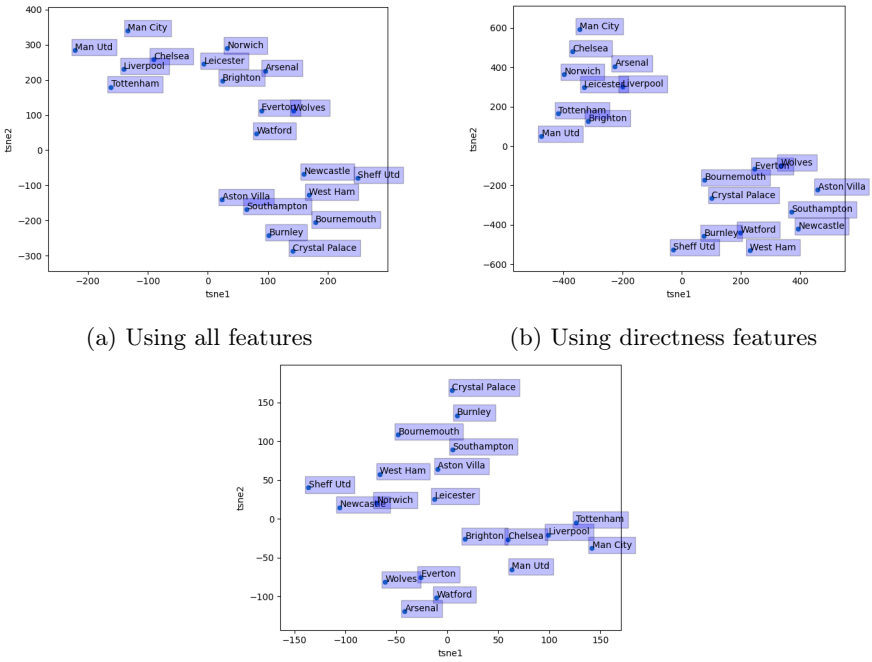
Teams who lose the ball often will achieve a lower score because there is a higher probability of being absorbed in the *move not successful* state. This is influenced by both the technical ability of a team as well as their behavior. While the influence of the technical ability is obvious, the influence of the behavior can be seen by means of an example. Consider a team that often attempts long balls. The probability of losing the ball when executing these passes will be higher because these tend to be more difficult. On the other hand, if these succeed, then the team is much closer to the shooting area and the likelihood of eventually arriving at a shot increases in this case. This is a trade-off that a team makes by adopting a specific playing style.

4 Use Cases

The previously defined features can be used to easily characterize and compare the intrinsic playing styles of teams. Using the event stream data of the 2019/20 English Premier League (EPL), we illustrate their use on three use cases: finding teams with similar playing styles, identifying mismatches in the relative efficiency of a team's style, and performing a more in-depth analysis of a teams' style.

4.1 Finding similar teams

Identifying teams with similar playing styles can be useful during match preparation, e.g., for measuring how similar your next opponent is to a team that you have played before and for scheduling pre-tournament friendlies against teams



(a) Using all features

(b) Using directness features

(c) Using positional features

Fig. 3: t-SNE visualization of all teams of the 2019/20 EPL using the three different options of combining the features. Teams which are visualized close to each other have a similar playing style.

that behave similarly to in-tournament opponents. We propose three different options of combining the features into playing style vectors. For each of these options, we visualize the vectors in a 2D-plane using t-SNE [17]. Similar data points will lie close to each other, allowing us to visually identify similar teams.

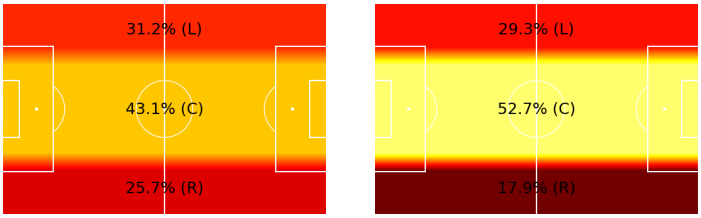
Option 1: Using all features. We can distinguish two clusters of teams when combining all 13 features defined in Section 3 into one playing style vector for each team (Figure 3a). One clear cluster is visible in the bottom right of the figure with teams like Sheffield United and Newcastle. These teams tend to have a more direct playing style and like to use their flanks. More possession-based teams like Manchester City can be found in the top part of the figure. Leicester City can be found the closest from all smaller teams to the big teams like Manchester City, Chelsea, and Liverpool. In the analyzed season, Leicester had a possession-based build up with quality players that were good at creating shooting opportunities. This resulted in them finishing in 5th place and playing the Europa League.

Option 2: Using all features regarding directness of playing style. We can distinguish two clear clusters of teams when only taking into account the features regarding the directness of play (Figure 3b). The cluster in the upper left corner contains teams that tend to have more possession, possibly because their styles of play focus on trying to maintain it. Manchester City is the extreme example, but teams such as Leicester and Brighton also preferred to maintain possession. Additionally, the majority of teams in this cluster have strong players. In contrast, weaker teams may be inclined to sit deep, absorb pressure and try to hit on the counter. The cluster in the bottom right of the figure contains teams with a more direct counter attacking style of play such as Aston Villa and Newcastle United. That season, Aston Villa preferred to play long balls to get the ball forward quickly, which was also made possible by the fast Jack Grealish. Under management of Steve Bruce, Newcastle United preferred to camp around their own goal, allowing the other team to take possession, and often attacked on the counter, which did not prove very fruitful for them.

Option 3: Using all features regarding the positional nature of teams. We can distinguish three clusters of similar teams when only taking into account the features regarding the locational preferences of teams (Figure 3c). The top of the figure contains teams such as Sheffield United and Crystal palace that tend to frequently use the flanks. Their ratio of inward/outward pointing sequences also indicates that they actively try to move the ball to the outside of the pitch. In contrast, the right side of the figure contains teams like Manchester City and Tottenham that tend to use the center of the field most often and also actively try to move the ball there. A last cluster of teams can be found in the bottom center of the figure containing Arsenal, Everton, Watford, and Wolverhampton. These teams divide their use of the field more equally. This could possibly be due to the teams changing tactics throughout the season as three out of these four teams (Arsenal, Everton, and Watford) changed managers mid-season.

4.2 Assessing mismatch in efficiency of the sides

Identifying possible mismatches in the efficiency of a team's playing style can be useful to create a game plan when playing against them, or to propose improvements when analyzing one's own style. To illustrate this, we inspect whether Bournemouth's expected usage of the sides and center of the field match up with their expected efficiency. Bournemouth use their flanks slightly more often than the league average (56.9% vs. 54.2%) and have a preference for the left over right flank with 21% more ball movements taking place on the former during their regular possession sequences (Figure 4a). However, when only considering possession sequences that end in a shot, there are 64% more actions taking place on their left vs. right side, which is much more than the 21% that would be expected if the sides were equally efficient (Figure 4b). Perhaps the team should have considered focusing even more on the left side when trying to attack. This could have been useful, as Bournemouth was relegated after the 2019/20 season.



(a) All sequences (b) Shot sequences

Fig. 4: Relative use of the left side, right side and central zone for Bournemouth according to their model when (a) all sequences and (b) only the sequences that end in a shot are taken into account.

4.3 In-depth analysis of playing style

Finally, we perform an in-depth analysis of the identified playing styles of Manchester City and Sheffield United.

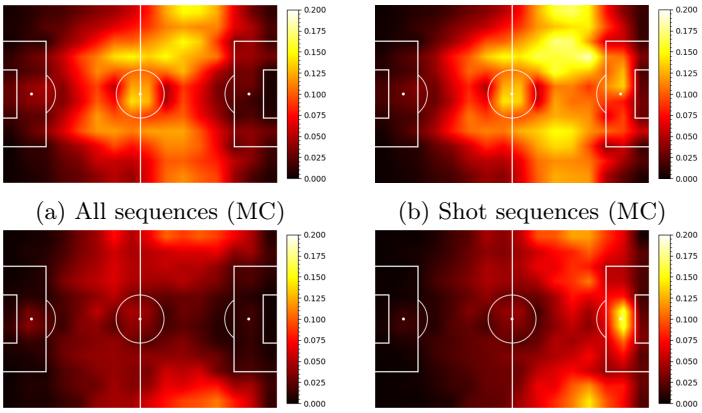
Regarding locational preference, City seems to prefer utilizing the half spaces during their build-up play, with a particular preference for the left half space (Figure 5). Their usage of the central zone is further emphasized when aggregating their usage of the left, right, and central zones (Figure 6). City uses the central zone more often (54.1%) than any other team with the league average being 45.8%. Consequently, they are also the team that is the least likely to use the sides, and when they do, they prefer the left side over the right side. On the other hand, Sheffield predominantly prefers the flanks and uses them more than any other team. This corresponds with their 3-5-2 formation where the outside center backs would overlap the wing backs to overload situations on the flanks.[5]

Regarding directness of play, City has the most elaborate buildup of all teams, with no other team having a higher average number of actions in sequences ending in a shot (14.2), or with lower probabilities of performing a long goal kick (4.3%) or using a long ball in the own half (0.9%). In contrast, Sheffield is one of the teams with the most direct playing style according to the model. Only Bournemouth has a lower average number of actions in possession sequences ending in a shot (7.2 vs. 8.0), and no team has a higher probability of performing a long goal kick (43.0%) or using a long ball in their own half (2.8%).

Regarding the ability to create shots, City has the highest probability of creating a shooting opportunity when possessing the ball in their own half (15.7%), with the league average being 9.0%. There is also no clear mismatch visible in their efficiency of the sides (Figure 6), which is not the case for all teams (see Section 4.2). Interestingly, City is the least likely team to generate a shot on a counterattack (1.7%) after regaining the ball in their own half. This emphasizes that they are extremely picky about when to launch a counterattack and do not risk losing the ball as they know how adept they are at generating shots with

[5] `https://themastermindsite.com/2020/08/29/overlapping-centre-backs-tactical-analysis/`

a patient build up. In contrast, Sheffield does not turn out to be good at creating shooting opportunities. When possessing the ball in their own half, they have the worst probability of generating a shot (5.6%). Traditional statistics for the 2019/20 season confirm the model's pessimistic view of their chance creation: they had the lowest average number of shots per game and only four teams scored fewer goals. Their ability to create successful counterattacks (2.2%) is also just below the league average of 2.3%. This suggests that the obtained 9th place during that season was generous based on their style of play and performance, and they were indeed relegated after the next season.



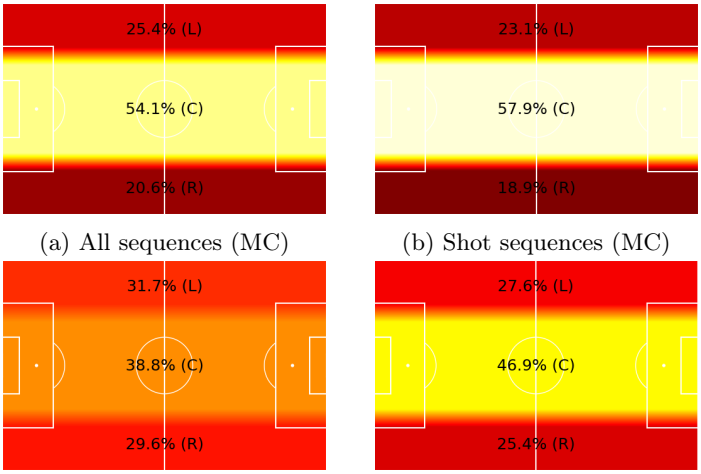
(a) All sequences (MC) (b) Shot sequences (MC)

(c) All sequences (SU) (d) Shot sequences (SU)

Fig. 5: Average number of visits to each location in a possession sequence for Manchester City (top row) and Sheffield United (bottom row) according to their model when all sequences (left column) and only the sequences that end in a shot (right column) are taken into account.

5 Related Work

Playing style analysis has already been approached from many different angles. Some works simply aim to retrieve the most common action patterns of a team by e.g., a combination of clustering and pattern mining [6,19] or inductive logic programming [18]. Other works adopt a more generalized view of playing style. For example, some apply clustering methods to the team's (ball) movements to identify the different behavior styles or prototypical actions that are used [1,7,8]. Other works aim to utilize compression methods such as Principal Component Analysis or non-negative matrix factorization to identify factors in the data of players that represents their playing style [2,5,9,11]. More recently, deep learning techniques have also been used to characterize a player's passing style [3].



(a) All sequences (MC) (b) Shot sequences (MC)

(c) All sequences (SU) (d) Shot sequences (SU)

Fig. 6: Relative use of the left side, right side and central zone for Manchester City (top row) and Sheffield United (bottom row) according to their model when all sequences (left column) and only the sequences that end in a shot (right column) are taken into account.

Most of these methods do not only try to analyze the playing style of teams, but also identify the different types of style. This is in contrast to the approach adopted in this paper. The identification of the different types of style according to the discussed techniques has two main disadvantages: 1) the identified categories are not always interpretable, and 2) the categorization ultimately depends on the choice of included features, and whether each feature is as significant to get a classification according to the intuitive and practical notion of playing style is often ignored. In contrast, we first define high-level indicators of playing style, break these down into a set of concrete features and then use analytical approaches and model checking techniques to derive values for these features. This has the advantage that the features are and remain interpretable from the start. Additionally, computing the feature values based on a model of the team's intrinsic offensive behavior yields values that are less influenced by rare actions.

Regarding the use of Markov models for soccer analytics tasks, there are many different applications with playing style analysis being the least researched. Rudd [14] first introduced the use of Markov models to the field of soccer analytics by using them to value player actions. The general idea was built upon by others [16,23]. In particular, the Expected Threat (xT) framework of Singh has been illustrated to be useful for analyzing the playing style of teams based on where teams generate threat from [16]. Our work encompasses this framework, as the same xT values can be computed using our proposed models. Peña [12] discusses how Markov models can be used to model possession sequences which yield faithful approximations of the distribution of passing sequences. Van Roy

et al. [20,21,22] use a Markov Decision Process instead to model possession sequences in which the policy reflects a team's historical action behavior. These models can be used to measure the effect of adjusting this behavior, to reason about defensive strategies, and to value a player's decision making. Markov models are also used in other sports. An example is the valuation of player actions in the National Hockey League using a model representing ice hockey games [13,15].

6 Conclusion

This paper proposed a novel approach to carry out playing style analysis. Instead of carrying out data analysis based directly on historical data, it first learns an intermediate team-specific Markov chain representing the offensive behavior of a team. These models can both capture the sequential patterns of a team's style as well as generalize over a team's historical behavior. That is, they capture slight variations on the playing style of teams, even when these are not explicitly observed in the limited amount of data. Additionally, we defined a number of features that characterize playing style and showed how analytical approaches and probabilistic model checking can be used to reason about each team's learned model to obtain values for these features. We illustrated our approach on teams in the 2019/20 English Premier League and showed how our approach can be used to 1) find teams with similar playing styles, 2) find inefficiencies in the playing style, and 3) perform an in-depth analysis of the playing style. The resulting insights can be used to guide coaches and managers when preparing for their next opponent or when scouting new players. Future work can propose more fine-grained models by including temporal information, additional states, and the intentions of actions, and by distinguishing between different action types.

Acknowledgements This work was supported by the Research Foundation – Flanders under EOS No. 30992574. We thank the RBFA Knowledge Centre for their valuable feedback.

References

1. Bialkowski, A., Lucey, P., Carr, P., Yue, Y., Sridharan, S., Matthews, I.: Identifying team style in soccer using formations learned from spatiotemporal tracking data. In: IEEE Int. Conference on Data Mining Workshop. pp. 9–14 (2014)
2. Castellano, J., Aguilar Pic, M.: Identification and preference of game styles in laliga associated with match outcomes. Int. journal of environmental research and public health **16**(24), 5090 (2019)
3. Cho, H., Ryu, H., Song, M.: Pass2vec: Analyzing soccer players' passing style using deep learning. Int. journal of sports science & coaching **17**(2), 355–365 (2021)
4. Decroos, T., Bransen, L., Van Haaren, J., Davis, J.: Actions speak louder than goals: Valuing player actions in soccer. In: Proceedings of the 25th ACM SIGKDD Int. Conference on Knowledge Discovery & Data Mining. p. 1851–1861 (2019)

5. Decroos, T., Davis, J.: Player vectors: Characterizing soccer players' playstyle from match event streams. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 569–584 (2019)
6. Decroos, T., Van Haaren, J., Davis, J.: Automatic discovery of tactics in spatio-temporal soccer match data. In: Proceedings of the 24th ACM SIGKDD Int. Conference on Knowledge Discovery & Data Mining. p. 223–232 (2018)
7. Decroos, T., Van Roy, M., Davis, J.: Soccermix: Representing soccer actions with mixture models. In: Proceedings of the 2020 Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 459–474 (2020)
8. Diquigiovanni, J., Scarpa, B.: Analysis of association football playing styles: An innovative method to cluster networks. Statistical modelling **19**(1), 28–54 (2019)
9. Fernandez-Navarro, J., Fradua, L., Zubillaga, A., Ford, P.R., McRobert, A.P.: Attacking and defensive styles of play in soccer: analysis of spanish and english elite teams. Journal of sports sciences **34**(24), 2195–2204 (2016)
10. Kwiatkowska, M., Norman, G., Parker, D.: PRISM 4.0: Verification of probabilistic real-time systems. In: Proc. 23rd Int. Conference on Computer Aided Verification (CAV'11). LNCS, vol. 6806, pp. 585–591, Springer (2011)
11. Lago-Peñas, C., Gómez-Ruano, M., Yang, G.: Styles of play in professional soccer: an approach of the chinese soccer super league. Int. journal of performance analysis in sport **17**(6), 1073–1084 (2017)
12. Peña, J.L.: A markovian model for association football possession and its outcomes. arXiv preprint arXiv:1403.7993 (2014)
13. Routley, K., Schulte, O.: A markov game model for valuing player actions in ice hockey. In: Uncertainty in Artificial Intelligence Conference. pp. 782–791 (2015)
14. Rudd, S.: A Framework for Tactical Analysis and Individual Offensive Production Assessment in Soccer Using Markov Chains. In: New England Symposium on Statistics in Sports (2011), `http://nessis.org/nessis11/rudd.pdf`
15. Schulte, O., Khademi, M., Gholami, S., Zhao, Z., Javan, M., Desaulniers, P.: A markov game model for valuing actions, locations, and team performance in ice hockey. Data Mining and Knowledge Discovery **31**(6), 1735–1757 (2017)
16. Singh, K.: Introducing expected threat. `https://karun.in/blog/expected-threat.html` (2019)
17. Van Der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**, 2579–2625 (2008)
18. Van Haaren, J., Dzyuba, V., Hannosset, S., Davis, J.: Automatically discovering offensive patterns in soccer match data. In: Proceedings of the 14th Int. Symposium on Intelligent Data Analysis. IDA, vol. 9385, pp. 286–297 (2015)
19. Van Haaren, J., Hannosset, S., Davis, J.: Strategy discovery in professional soccer match data. In: KDD-16 Workshop on Large-Scale Sports Analytics. pp. 1–4 (2016)
20. Van Roy, M., Robberechts, P., Yang, W.C., De Raedt, L., Davis, J.: Learning a markov model for evaluating soccer decision making. In: RL4RealLife workshop at ICML (2021)
21. Van Roy, M., Robberechts, P., Yang, W.C., De Raedt, L., Davis, J.: Leaving goals on the pitch: Evaluating decision making in soccer. In: Proceedings of the 15th Annual MIT Sloan Sports Analytics Conference (2021)
22. Van Roy, M., Yang, W.C., De Raedt, L., Davis, J.: Analyzing learned markov decision processes using model checking for providing tactical advice in professional soccer. In: AI for Sports Analytics (AISA) workshop at IJCAI (2021)
23. Yam, D.: Attacking contributions: Markov models for football. `https://statsbomb.com/2019/02/attacking-contributions-markov-models-for-football/` (2019)