Analyzing passing sequences for the prediction of goal-scoring opportunities

Conor McCarthy[1], Panagiotis Tampakis[1], Marco Chiarandini[1], Morten Bredsgaard Randers[2,3], Stefan Jänicke[1], and Arthur Zimek[1]

[1] Department of Mathematics and Computer Science, University of Southern Denmark, Denmark
conor099@hotmail.com,{ptampakis,marco,stjaenicke,zimek}@imada.sdu.dk
[2] Department of Sports Science and Clinical Biomechanics, University of Southern Denmark, Denmark
mranders@health.sdu.dk
[3] School of Sport Sciences, Faculty of Health Sciences, UiT The Arctic University of Norway, Tromsø, Norway

Abstract. Over the last years, more and more sport related data are being collected, stored, and analyzed to give valuable insights. Football is no exception to this trend. An important way of identifying a team's "style" of play is through analyzing passing sequences. However, passing sequences either concentrate on the specific players involved or the structure of passes and ignore where these sequences took place. In this paper, we focus on identifying frequent passing zone subsequences that lead to created or conceded goal scoring opportunities. We partition the pitch into a set of disjoint zones and apply sequential pattern mining. Our experimental study on the 2020/21 Danish Superliga season shows that our method is able to predict goal scoring opportunities better than random subsequences that occurred, in median, 99.5% of the cases.

1 Introduction

During the last years, the analysis of data from professional football has attracted a lot of attention. A hot topic in the football world is a team's "style" of play. There are numerous styles that teams commonly have, including: possession, high pressing, and direct play. How teams try to score or concede goals is a defining factor in their style of play, as, ultimately, the goal of a game of football is to try to score more goals than the other team. Towards this direction, there have been some efforts for analysing offensive and defensive strategies [3, 18, 13]. However, most of them focus on plain statistical analysis, which could ignore hidden patterns that might exist in the data and that could be identified by more knowledge discovery techniques on tracking data, such as clustering [17, 12] and outlier detection [15, 21]. A popular way of analysing the style of play of a team is through its passing sequences [6, 1], where a passing network gets constructed, which is a graph where the nodes represent the players and the edges the passes between players. Nevertheless, this approach concentrates on

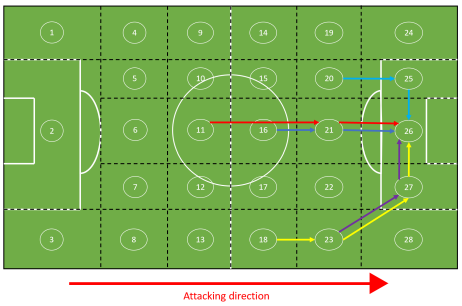


Fig. 1: Visualisation of Brøndby IF's top 3 most frequent zone subsequences that lead to a goal-scoring opportunities.

the identity of the players involved in passing sequences rather than on roles [14] or on the structure of the passes risks to capture only a partial representation of a team's play style. Flow motifs [9], which are sequences of three consecutive, uninterrupted passes involving a maximum of four distinct players, where labels represent distinct players without identity, try to overcome this limitation by focusing on the structure of a passing sequence. To exemplify, let us consider a passing sequence 2-4-2-4, where player 2 passes to player 4, player 4 passes back to player 2 and player 2 back to player 4. The corresponding flow motif is ABAB. Then, 4-2-4-5 (the next consecutive 3 passes) turns into ABAC. On the other hand, flow motifs disregard the area on the pitch where the passing sequence took place, which might give us significant insight.

Motivated by these previous works, we also focus on identifying frequent passing sequences that lead to created or conceded goal scoring opportunities. However, rather than considering individual players we consider areas on the pitch from where the passes occur. To achieve this, we partition the pitch into a set of disjoint zones and utilize these zones to discover *frequent zone subsequences* that lead to a created or conceded goal scoring opportunity, as illustrated in Figure 1. The rest of the paper is organized as follows, in Section 2 we formally define the problem, in Section 3 we provide details about our methodology and in Section 4 we present the findings of our study. Finally, in Section 6 we conclude and discuss about future directions of our work.

2 Problem Definition

Definition 1. *(**Tracking Data**) The tracking data T of a game consist of a sequence of semantically enriched timestamped locations $\langle p_1, p_2, \ldots, p_N \rangle$, where N is the number of samples. Each data point p_i, $i \in [1, N]$ consists of a tuple $p_i = (x_i, y_i, z_i, t_i, tp_i)$, where x_i, y_i, z_i designate the position of the ball in the pitch, t_i the timestamp, and tp_i the team in possession of the ball at time t_i.*

Definition 2. *(**Tracking Subsequence**) A tracking subsequence $T_{i,j}$, $i < j$, is a subsequence $\langle p_i, \ldots, p_j \rangle$ of T that contains all semantically enriched timestamped locations of the ball between t_i and t_j.*

Definition 3. *(**Event Data**) Each game consists of a set of events $E = \{e_1, e_2, \ldots, e_M\}$. Each event e_j for $j \in [1, M]$ consists of a tuple $e_j = (x_j, y_j, t_j, et_j)$, where x_j, y_j denote the x and y coordinates, t_j the timestamp, and $et_j \in ET$ the event type.*

The set of event types ET contains different kinds of events, such as pass, foul or interception. These events can be grouped into two large groups, i.e., events that change the possession of the ball, called CPE, e.g., an interception or a successful tackle, and events that do not change the possession, called $nCPE$. Clearly, $CPE \cap nCPE = \emptyset$. Thus $ET = CPE \cup nCPE$. We assume that the tracking and event data have been aligned (see later for details), that is, the timestamp and coordinates of every event $e_j \in E$ have been matched with exactly one data point from the tracking data T. We break the large sequence of tracking and event data that correspond to an entire game into smaller portions, each portion corresponding to a ball possession by a team, by utilizing CPE.

Definition 4. *(**Possession Subsequences**) The collection of ball possessions for a specific team is denoted by $PS = \{ps_1, ps_2, \ldots, ps_Q\}$. Each ps_k corresponds to a tracking subsequence T_{i_k,j_k} of T with $et_{i_k}, et_{j_k} \in CPE$ and $et_l \in nCPE$ for all $l \in [i_k + 1, j_k - 1]$.*

We are interested not in all possessions but only in those leading to a goal scoring opportunity. Such events are possession for a team ending with them scoring a goal, winning a penalty, or taking a shot on goal that did not end in a goal.

Definition 5. *(**Goal Scoring Opportunity Subsequences**) The collection of ball possessions that lead to a goal scoring opportunity is denoted by $GSO = \{gso_1, gso_2, \ldots, gso_R\}$. GSO is a subset of PS, $GSO \subseteq PS$, and for each gso_k, $k \in [1, R], R < Q$, it is $et_{j_k} \in \{Goal, Penalty, ShotOnGoal\}$.*

Several knowledge extraction techniques, such as sequential pattern mining, would not benefit by the exact position of the ball at a specific timestamp. For this reason, we decided to ignore the temporal dimension and consider only the sequence of ball positions. Furthermore, we chose to abstract the spatial dimension at a higher level by partitioning the pitch in a set of zones.

Definition 6. *(**Pitch Partitioning**) A pitch can be considered to consist of a set of zones $P = \{z_1, z_2, \ldots, z_S\}$, where each $z_i \in P$ is a rectangle defined by $(zid_i, lx_i, ly_i, hx_i, hy_i)$, with zid_i being the zone identifiers and lx_i, ly_i and hx_i, hy_i being the coordinates of the lower and higher corners, respectively. It holds that $\cap_{i \in [1,S]} z_i = \emptyset$ and $\cup_{i \in [1,S]} z_i$ covers the whole pitch.*

Now, by spatially overlapping each $ps_k \in gso_m \in GSO$ with P, we can remove the exact spatial position and replace it with the zone id. Furthermore,

we can remove the exact timestamp and keep only the sequence of zones. Finally, we can eliminate successive duplicate zone appearances. By doing so, we result in having zone sequences instead of opportunity possessions.

Definition 7. *(**Goal Scoring Opportunity Zone Sequence**) Formally, a zone sequence $zs \in ZS$ is defined as $zs = \{zid_1, zid_2, \ldots, zid_{|zs|}\}$, where $zid_i \neq zid_j \forall i, j \in [1, |zs|]$. It is obvious that $\forall gso_m \in GSO$ we get a $zs_m \in ZS$.*

Each goal scoring opportunity zone sequence $zs \in ZS$ is a string of identifiers (e.g., integer numbers) for which we distinguish substrings and subsequences.

Definition 8. *(**Zone Substring**) A zone substring $zs_{i,j}$ of $zs = \langle zid_1, zid_2, \ldots, zid_\ell \rangle$ is defined as $\langle zid_i, zid_{i+1}, \ldots, zid_j \rangle$, that is, the set of consecutive elements from zs between position $i \geq 1$ and position $j \leq \ell$, where $i \leq j$.*

Definition 9. *(**Zone Subsequence**) A zone subsequence zs' consists of $\langle zs_{i,j}, zs_{k,l}, \ldots zs_{m,n} \rangle$, with $i \leq j < k \leq l < m \leq n$, can be defined as a set of zone substrings contained in zs.*

Definition 10. *(**Support**) For a collection of zone sequences of goal scoring opportunities, $ZS = \{zs_1, zs_2, \ldots, zs_R\}$, the support of a zone subsequence zs' is defined as* $\text{support}(zs', ZS) = |\{zs \in ZS \mid zs' \text{ is a subsequence of } zs\}|/R$.

We can now formalize our problem as follows.

Definition 11. *(**Frequent Zone Subsequence Discovery**) Given a set T of tracking data, a set E of event data, a pitch partitioning P and a minimum support threshold τ, our task is to retrieve from the collection of zone sequences of goal scoring opportunities, ZS, the set ZS^* of zones subsequences with support above the threshold τ, that is, to determine $ZS^* = \{zs' \mid \exists\ zs \in ZS, zs' \text{ is subsequence of } zs \wedge support(zs', ZS) \geq \tau\}$.*

3 Methodology

In Figure 2 we summarize the methodology that we used to tackle the problem defined in Section 2. Initially, we perform a preprocessing step to align the event and tracking data. Subsequently, we utilize the enriched tracking data to extract the goal scoring opportunity subsequences *GSO*. The pitch partitioning step can be performed either successively or in parallel with the *GSO* extraction step, depending on whether the partitioning is data-driven or not. Next, the *GSO*s are compressed into Zone Sequences. Finally, the Frequent Zone Subsequences are discovered by preforming Sequential Pattern Mining on the Zone Sequences.

3.1 Tracking Data

The actual tracking data utilized in this paper were provided by ChyronHego, which used the TRACAB Image Tracking System [2] to track the location of

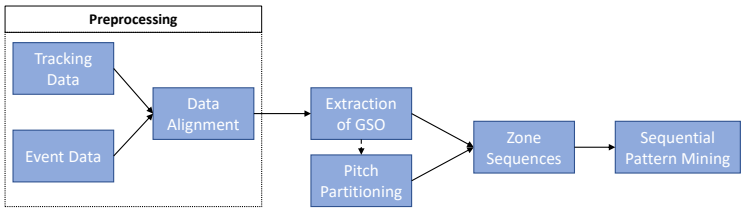


Fig. 2: Overview of the methodology.

players, referees, and the ball (at a frequency of 25Hz) during matches for the Danish Superliga during the 2020/21 season. In our case we utilized only the location of the ball. Each tuple of the dataset consists of (1) x, (2) y, (3) z, (4) the speed, (5) the team in possession and (6) the ball status (ball in play or not). For our analysis, we used only the attributes that are defined in Definition 1 and the tuples where the ball is in play.

3.2 Event Data

The event data for each match are created by the company Opta. It should be noted that the Opta event data are not automated but created manually by people who are assigned to watch a game live. Each tuple of the event data consisting of (1) an *id*, (2) a *type_id*, (3) *period_id* (1st or 2nd half), (4) t (the timestamp), (5) x, (6) y (7) *player_id*, (8) *team_id*, (9) *outcome*, and (10) *qualifiers* (additional detail about an event). At this point, we should mention that every shot taken throughout a match has a qualifier that states the expected goal value (xG) for that shot. The expected goal value, xG, is a measure of the quality of a scoring opportunity based on the probability of a shot being performed from a particular position during a particular passage of play. The model used by Opta takes many different variables into account when predicting this probability of a shot leading to a goal [4]. It uses logistic regression with the response variable being whether a shot resulted in a goal or not. From these values, it is possible to get a decent idea as to how likely a shooting opportunity is to end in a goal.

3.3 Data Alignment

As already mentioned, the event data were manually annotated and for this reason there might exist inconsistencies between the event data and the tracking data. For the analysis of these data, it was necessary to align these two datasets and in order to achieve this we utilized the work done by Kuzmicki [8], where the events are divided into four categories (sub-problems): (1) state-changing events, (2) pass events, (3) duels, and (4) recovery events. The state-changing events are those when the ball goes out of play and becomes inactive, such as throw-ins and goalkicks. Pass events consist of passes or clearances, while duels are situations in which two players on opposing teams compete for the ball such

as in headers or tackles. Finally, recovery events consist of events where the ball is recovered by a team, such as interceptions and goalkeeper claims. The method from [8] first aligns state-changing events, then looks at the times between state-changing events, where the ball is in play, and aligns the pass events. Finally, it aligns duels and then recovery events.

To align the state-changing events, the times in the tracking data when the status of the ball goes from alive to dead were considered and aligned with the corresponding state-changing events that were within two seconds, according to the timestamps in the event data. The event-types that are aligned in this manner are, throw-ins and goalkicks (outs), corners, offsides, fouls, goals, game/half starting, and game/half ending. Once the state-changing events are aligned, to align the passes, two sequences were extracted between every state-changing event when the ball is in play, one from the tracking and one from the event. From the event data, the sequence of pass performers between the state-changing events were looked at, and from the tracking data, the sequence of 'ball possessors' were extracted, where a 'ball possessor' is a person that is within 1.5 metres of the ball during any frame between the two state-changing events, as defined by Kuzmicki. These two sequences are then aligned using the Needleman-Wunsch algorithm [10]. The event-types that are aligned in this manner are, passes, offside passes, clearances, shot misses, shots that hit the post, saves by goalkeepers, and ball touches. Some of these events obviously aren't actually passes, but they can be aligned using the same principle. After the state-changing events and passes have been aligned, the events where duels occur are aligned. For duel events, there are two events in the event data, one for each player involved. To align these events, the distances between both players and the ball are calculated, and the frame where the minimum sum of these distances occurs is aligned with the two events. Finally, for the alignment of the recovery events, the timeframe between events where these events can occur, is much smaller, due to the other events already previously being aligned. The recovery events are aligned by looking at these timeframes, and aligning the instant that the ball comes within 1.5 metres of the player doing the event. Some events get misaligned in the process of the sequence alignment, this can happen for a number of reasons, the event data assigned an event to the wrong player, an event was recorded in the event data that never occurred, or an event occurred but was never recorded in the event data. Kuzmicki added an enrichment process to the alignment where he corrected some of the mismatched events and also added other event-types. The correction of some of the events is based on the assumption that the tracking data is more accurate than the event data, as the tracking is automated and the event is manual.

3.4 Extraction of Goal Scoring Opportunities

To extract the goal scoring opportunities we follow a slightly different approach than the one defined in Definition 5. More specifically, we rely on a previous work [5]. Accordingly, an expected goal value, xG, greater than or equal to 0.33

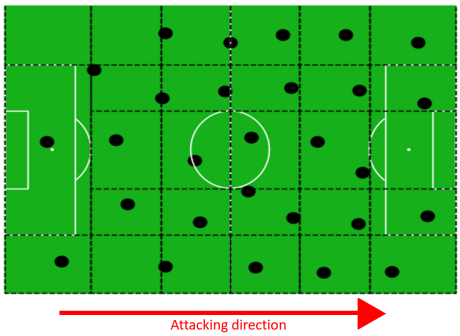


Fig. 3: Selected pitch partitioning, overlapped with the centroids produced by k-means.

constitutes a "big chance" for a team. Hence, we considered as goal scoring opportunities only sequences with xG above this threshold.

Goals that occur from set pieces, i.e., corners, free kicks, do not produce very interesting sequences, as they contain either just a cross into the box or a shot, and the team is not really in physical control of the ball when it is in the air after being crossed. Therefore, even if it is estimated that approximately 30-40% of all goals in professional football are from set pieces [20], nevertheless, we decided to exclude all set pieces from our analysis.

3.5 Pitch Partitioning

Once the goal-scoring opportunity sequences have been extracted, the next step is to see what areas of the pitch the ball passes through during these sequences. Different ways of partitioning the pitch have been proposed [7, 19] and any of them could be used here. For this work, we decided to use a similar method to the one proposed in [16] because it includes the "half spaces" that are of particular interest to football coaches. However, we changed the defensive third into 3 zones instead of 5, as there are substantially fewer passes in these areas, thus, yielding 28 instead of 30 zones. We also decided to flip the numbers so that a team's defensive areas are the lower numbers and attacking areas are the higher, always. The selected pitch partitioning can be seen in Figure 3. However, other disjoint pitch partitioning solutions can be applied seamlessly, since this is orthogonal to our problem.

An alternative approach was to use the data in the goal-scoring opportunity sequences to infer the areas of the pitch from where passes are most frequently performed. We pooled the coordinates of every pass from all teams' goal-scoring opportunity sequences to get a generalised idea of all pass locations. We actually utilized this approach to verify the selected partitioning by applying the k-means clustering method (with k=28). The results illustrated in Figure 3 showed that, even though the partitioning selected above and the zones produced by the k-

means approach are not perfectly aligned, they have a striking similarity, since most of the zones only contain one cluster centroid.

3.6 Sequential Pattern Mining

Finally, we apply a Sequential Pattern Mining algorithm to extract the most Frequent Zone Subsequences. For this purpose, we utilized the PrefixSpan sequential pattern mining algorithm [11]. The algorithm is a depth first search (DFS) algorithm that mines for *subsequences* instead of *substrings*. As defined in Section 2, the actual difference between subsequences and substrings is that a substring is a set of consecutive zones, while a subsequence is a set similar to a substring, with the difference that it allows for zone "gaps". The reason for using subsequences instead of substrings is that the process of identifying frequent subsequences is less constrained than the process of identifying frequent substrings, hence it will produce more frequent patterns to analyse, yielding more robust results.

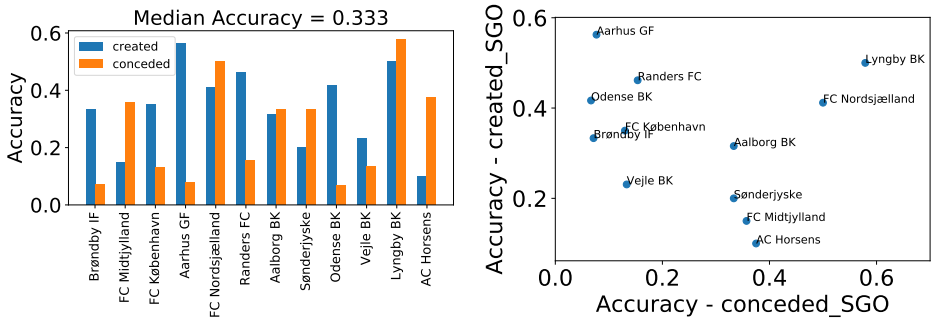
4 Experimental Study

Our experimental assessment is as follows. Initially, we apply our methodology and extract the Frequent Zone Subsequences, both for created and conceded goal scoring opportunities. Then, we calculate an accuracy measure to assess how accurate the sequential patterns found are for predicting goal-scoring opportunities for each team. Finally, this accuracy measure is compared to frequently occurring zone subsequences that are randomly extracted from the dataset. The length of these subsequences is equal to the length of the discovered frequent zone subsequences that lead to a goal scoring opportunity. These subsequences represent those that are commonly found during a match, and hence comparing their accuracy against the one of the discovered Frequent Zone Subsequences unveils how unlikely it is for a random subsequence to achieve better accuracy than the discovered Frequent Zone Subsequences.

We split the data into a training set and a test set. Thus, for each team's zone sequences of goal-scoring opportunities throughout the whole season, a training set ZS_{train} is extracted consisting of 70% of these sequences. In this training set, the frequent zone subsequences ZS^* are extracted for each team. The remaining 30% are then used as the test set ZS_{test}, which is used to test if the frequent zone subsequences ZS^* from the training set ZS_{train} actually appear in the unseen test set zone sequences as goal-scoring opportunities.

We measure the accuracy of ZS^* as the percentage of the number of sequences from ZS_{test} that contain at least one sequence from ZS^* as subsequence, over the total number of sequences in the test set. More formally, for a team the accuracy of its ZS^* can be defined as

$$\text{Accuracy} = \frac{|\{zs \in ZS_{\text{test}} \mid \exists\, \widehat{zs} \in ZS^*, \widehat{zs} \text{ is subsequence of } zs\}|}{|ZS_{\text{test}}|}. \quad (1)$$



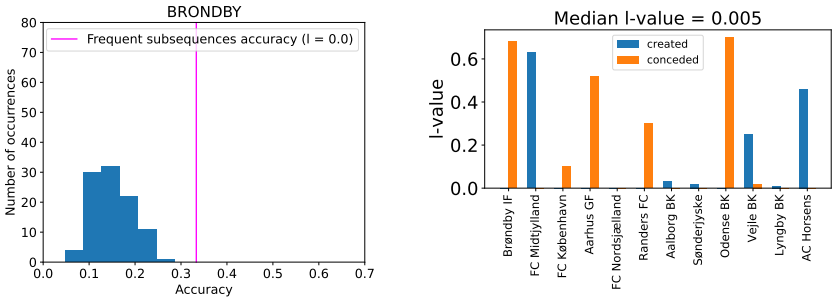
(a) (b)

Fig. 4: Accuracy of created vs conceded *GSO* (a) bar chart and (b) scatter plot.

The accuracy results for both created and conceded *GSO*s are depicted in Figure 4. The x-axis of Figure 4(a) shows the teams sorted from left to right according to their final ranking, while the y-axis reports the accuracy for both created and conceded goals. The median overall accuracy is 0.333, but it is hard to interpret if this is a "good" accuracy.

Figure 4(b) illustrates how each team performs in terms of accuracy, both for created and conceded *GSO*s. For example, we can clearly see that Lyngby is the dominant team in terms of conceded *GSO* accuracy and in terms of created *GSO* accuracy. This can be interpreted as an indication that this specific team plays in a predictable way both offensively and defensively, which led to poor performance as the final ranking of this team was second to last. On the other hand, Vejle, which seems to be one of the most unpredictable teams finished third from the bottom. A possible explanation to these opposite results for the two teams might be the different number of sequences that led to a *GSO* in the test set, which affects the denominator of Equation 1.

To deal with this, we tried to measure how much "better", in terms of accuracy, ZS^* performs in the test set, in comparison to a set of random subsequences ZS^* drawn from each team's top 500 most frequent zone subsequences that occurred throughout the season extracted from all possession sequences PS (hence, irrespective of whether they ended in a *GSO*). We construct the empirical distribution of the accuracy of 100 random subsequences ZS^* as follows. For each team, we repeat 100 times the sampling of a set of subsequences ZS^* of the same size as ZS^* from the 500 subsequences. Let B be the set of the sampled ZS^*. Then, for each random sample $ZS^* \in B$, the accuracy on the test set ZS_{test} is calculated. From these 100 accuracy values we derive the empirical distribution as illustrated in Figure 5. The likelihood of the accuracy of ZS^* is then the fraction of the samples from B that have a lower accuracy. A value of the likelihood close to zero indicates that the frequent zone subsequence is exceptional.



(a) (b)

Fig. 5: (a) Distribution of accuracy for Brondby's random subsequences (the pink line is the accuracy of ZS^*) and (b) likelihood of the accuracy of ZS^*.

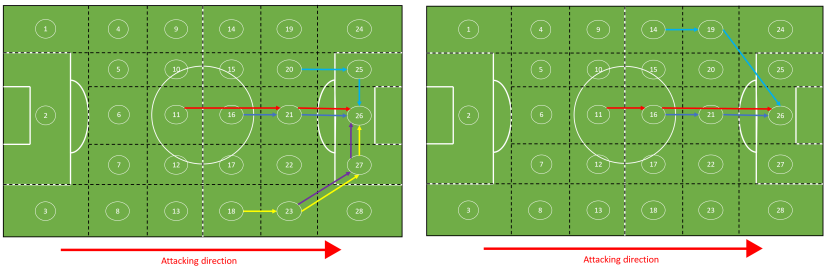
Figure 5(a), shows the distribution of the accuracy of random sequences along with the accuracy achieved (pink line) by the discovered frequent zone subsequences that lead to a GSO. Figure 5(b), depicts the likelihood value for each team both for created and conceded GSOs. We can see that most of the teams achieve a likelihood value close to zero. In addition, we can observe that the median likelihood value is 0.005=0.5%, indicating that the frequent zone subsequences from the proposed methodology lead to exceptionally high accuracy. Regarding Lyngby and Vejle, we can observe that Lyngby's extracted frequent zone subsequences are significant, since it has an l-value close to zero both for conceded and created opportunities. On the other hand, Vejle has a high value for created opportunities, which means that the extracted frequent zone subsequences for created opportunities are not so "trustworthy" and we should not base our analysis on them.

5 Style of play for the top-2 teams

In this section we focus on the top-2 teams individually and use their frequent zonal subsequences from goal-scoring opportunity sequences, as well as their frequent pass locations during these sequences, to get an idea as to what "style" of play each team has.

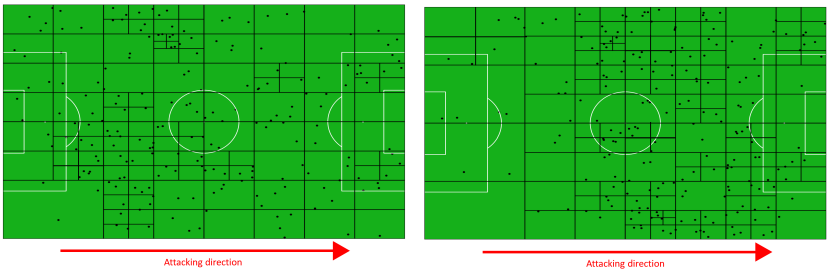
The winners of the league during the Danish Superliga 2020/21 season were Brøndby IF (BIF). Over the course of the season, they created the most goal-scoring opportunities, 67, and they conceded the second-least, 43, for teams in the Championship round. From the analysis of the significance of their frequent subsequences on these chances, presented in the previous section, it was seen that their observed accuracy was significantly better than their random accuracies for their created opportunities, but not for their conceded opportunities.

This suggests that the top-3 frequent subsequences from created goal-scoring opportunities found for BIF over the course of the season were highly significant



(a) (b)

Fig. 6: Top 3 most frequent zone subsequences that lead to a goal-scoring opportunities of (a) Brøndby IF and (b) Midtjylland.



(a) (b)

Fig. 7: Quadtree on passes made by (a) Brøndby IF and (b) Midtjylland.

in explaining how they created opportunities. The top 3 frequent subsequences found in all these goal-scoring opportunity sequences can be seen in Figure 6(a). It should be noted that because the sequences are subsequences, which allow gaps, the visualisation does not necessarily perfectly show the exact path for each subsequence. In Figure 7(a), the areas where most passes occurred during goal-scoring opportunities are visualised, using the Quadtrees method of splitting the pitch to show the most dense areas.

Combining the two figures showing the frequent subsequences and the pass densities, it appears that BIF tend to create a lot of their opportunities down the middle of the pitch with passes starting from their own half. They also appear to make a lot of use of the "half-spaces" 25 and 27, with a lot of passes occurring in both. Interestingly, there are 2 frequent subsequences attacking down the right side of the pitch, but not an abundance of passes actually occur here, perhaps suggesting that BIF play passes through this part of the pitch a lot, but don't actually make that many passes in them. The high density of passes just inside their own half on the right also indicate this. Indeed, BIF had a lot of shots after fast breaks, which would explain the sequences from the right, which indicate

that BIF counterattacked down this side a lot during the season. Overall, the majority of BIF's passes during the sequences occur in their own half of the pitch, which could suggest that they are a patient team that enjoy building attacks from the back, but are also capable of quick counterattacks.

The second-placed team over the season was FC Midtjylland (FCM). During the season, they created the third most goal-scoring opportunities, 62, and they conceded the joint second-least, 43, for teams in the Championship round. Analysing the significance of FCM's frequent subsequences from goal-scoring opportunity sequences, their subsequences for created chances were not found to be significant, but the subsequences for their conceded were not. The fact that their observed accuracy was not found to be significant could indicate that FCM are a more unpredictable team than others when creating chances, as they have many different areas of the pitch that they look to attack from. FCM's top 3 frequent zonal subsequences from goal-scoring opportunities can be seen in Figure 6(b).

As FCM's frequent subsequences were not found to be significant to how they create chances, there should be less emphasis on looking at these to define how they try to create goal-scoring opportunities. From the Quadtrees split of their passes, as it can be seen in Figure 7(b), many of their passes occur down the two sides of the pitch, on the wings. FCM's passes are also located much more in the opponent's half of the pitch compared to Brøndby IF, this could indicate that FCM are a much higher pressing team, gaining success from creating chances when winning the ball in the opponent's half. Overall, FCM seem to be a team that have a lot of variation in how they attack, with their frequent subsequences not being significant, and they appear to emphasise intricate passing in the opponent's half of the pitch in order to create chances. Indeed, FCM has a very dynamic attack with players constantly changing positions, and that they tend to dominate matches with a lot of possession and high pressure.

6 Conclusions and Future Work

We focused on identifying frequent passing zone subsequences that lead to created or conceded goal scoring opportunities. We proposed a methodology, consisting of (1) a preprocessing step that aligns event and tracking data, (2) the extraction of goal scoring opportunities, (3) the partitioning of the pitch into a set of zones, (4) the extraction of zone sequences, and (5) the discovery of frequent zone subsequences. The results indicate that our method is able to perform better than random subsequences that occurred, in median, 99.5% of the cases.

For the future, we plan to cross validate the results on different settings of training and test sets, either by using k-fold cross validation, or by using the first round as a training set and trying to predict the second round. Moreover, we plan to perform some sensitivity analysis with respect to the value of xG. Furthermore, we would like to work on a good methodology of data-driven pitch partitioning. Finally, we plan to use a spatial sequence similarity measure for the calculation of accuracy, instead of the containment presented in Equation 1.

References

1. Barbosa, A., Ribeiro, P., Dutra, I.: Similarity of football players using passing sequences. In: Machine Learning and Data Mining for Sports Analytics - 8th International Workshop, MLSA 2021, Virtual Event, September 13, 2021, Revised Selected Papers. Communications in Computer and Information Science, vol. 1571, pp. 51–61. Springer (2021)
2. ChryonHego: TRACAB optical tracking product information sheet. Tech. rep., ChryronHego (2019), https://chyronhego.com/wp-content/uploads/2019/01/TRACAB-PI-sheet.pdf, uRL: https://chyronhego.com/wp-content/uploads/2019/01/TRACAB-PI-sheet.pdf
3. 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)
4. Gregory, S.: Expected Goals in Context (2017), https://www.statsperform.com/resource/expected-goals-in-context/
5. Hernanz, J.: How good is Driblab's Expected Goals (xG) model? (2021), https://www.driblab.com/analysis-team/how-good-is-driblabs-expected-goals-xg-model/
6. Hughes, M., Franks, I.: Analysis of passing sequences, shots and goals in soccer. Journal of Sports Sciences **23**(5), 509–514 (2005)
7. Kim, J., James, N., Parmar, N., Ali, B., Vučković, G.: The Attacking Process in Football: A Taxonomy for Classifying How Teams Create Goal Scoring Opportunities Using a Case Study of Crystal Palace FC. Frontiers in Psychology **10**, 1–8 (2019)
8. Kuźmicki, P.: Synchronizaton, enrichment and visualizaton of football data. Master's thesis, University of Southern Denmark (SDU) (2020)
9. Malqui, J.L.S., Romero, N.M.L., Garcia, R., Alemdar, H., Comba, J.L.: How do soccer teams coordinate consecutive passes? A visual analytics system for analysing the complexity of passing sequences using soccer flow motifs. Computers & Graphics **84**, 122–133 (2019)
10. Needleman, S.B., Wunsch, C.D.: A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology **48**(3), 443–453 (1970)
11. Pei, J., Han, J., Mortazavi-Asl, B., Pinto, H., Chen, Q., Dayal, U., Hsu, M.C.: PrefixSpan,: mining sequential patterns efficiently by prefix-projected pattern growth. In: Proceedings 17th International Conference on Data Engineering. pp. 215–224 (2001)
12. Pelekis, N., Tampakis, P., Vodas, M., Panagiotakis, C., Theodoridis, Y.: In-dbms sampling-based sub-trajectory clustering. In: Proceedings of the 20th International Conference on Extending Database Technology, EDBT 2017, Venice, Italy, March 21-24, 2017. pp. 632–643. OpenProceedings.org (2017)
13. Rahimian, P., Toka, L.: Inferring the strategy of offensive and defensive play in soccer with inverse reinforcement learning. In: Machine Learning and Data Mining for Sports Analytics - 8th International Workshop, MLSA 2021, Virtual Event, September 13, 2021, Revised Selected Papers. Communications in Computer and Information Science, vol. 1571, pp. 26–38. Springer (2021)
14. Sattari, A., Johansson, U., Wilderoth, E., Jakupovic, J., Larsson-Green, P.: The interpretable representation of football player roles based on passing/receiving patterns. In: Machine Learning and Data Mining for Sports Analytics - 8th International Workshop, MLSA 2021, Virtual Event, September 13, 2021, Revised Selected

Papers. Communications in Computer and Information Science, vol. 1571, pp. 62–76. Springer (2021)
15. Schubert, E., Zimek, A., Kriegel, H.: Local outlier detection reconsidered: a generalized view on locality with applications to spatial, video, and network outlier detection. Data Min. Knowl. Discov. **28**(1), 190–237 (2014)
16. Seymour, D.: Tactical Theory: Using the half-spaces to progress the ball (2020), https://totalfootballanalysis.com/article/tactical-theory-using-half-spaces-progress-ball-tactical-analysis-tactics
17. Tampakis, P., Pelekis, N., Doulkeridis, C., Theodoridis, Y.: Scalable distributed subtrajectory clustering. In: 2019 IEEE International Conference on Big Data (IEEE BigData), Los Angeles, CA, USA, December 9-12, 2019. pp. 950–959. IEEE (2019)
18. Tenga, A., Holme, I., Ronglan, L.T., Bahr, R.: Effect of playing tactics on achieving score-box possessions in a random series of team possessions from norwegian professional soccer matches. Journal of Sports Sciences **28**(3), 245–255 (2010)
19. Tianbiao, L., Andreas, H.: Apriori-based diagnostical analysis of passings in the football game. In: 2016 IEEE International Conference on Big Data Analysis (ICBDA). pp. 1–4 (2016)
20. Yiannakos, A., Armatas, V.: Evaluation of the goal scoring patterns in European Championship in Portugal 2004. International Journal of Performance Analysis in Sport **6**, 178–188 (2006)
21. Zimek, A., Filzmoser, P.: There and back again: Outlier detection between statistical reasoning and data mining algorithms. WIREs Data Mining Knowl. Discov. **8**(6) (2018). https://doi.org/10.1002/widm.1280