15TH ANNUAL
MIT SLOAN
SPORTS ANALYTICS CONFERENCE
APRIL 8-9, 2021 | VIRTUAL
presented by ESPN 42 ANALYTICS

Routine Inspection: A playbook for corner kicks

Laurie Shaw, Harvard University and Sudarshan Gopaladesikan, SL Benfica

1. Introduction

A corner kick is awarded in soccer when the ball goes out of play over the goal line (but not into the goal) having last been touched by a member of the defending team. The attacking team restarts play by kicking the ball from the corner of the field closest to where the ball went out. Often, teams will attempt to create a goal scoring opportunity by crossing the ball directly into their opponent's penalty area, aiming for one of several attacking players maneuvering in the penalty area.

On average, teams win approximately five corner kicks per game. This provides each team with five opportunities to execute a corner strategy that they have rehearsed in training. Although the average conversion rate, the proportion of corners that directly produce a goal, is very low (around 2% in most elite leagues[1]), some teams have achieved much higher conversion rates, attributed to careful planning and meticulous execution of corner strategies [1]. FC Midtjylland are a well-known example, dedicating up to four training sessions per week to practicing corners [2]. Midtjylland won the Danish Superliga title in 2014-15 (and again in 2017-18) scoring an unusually high proportion of goals from corners and other set-piece situations**.**

How do teams like Midtjylland successfully convert corner kicks into goals? Previous work has focused on general metrics for defining offensive corner strategy, such as the ball delivery trajectory (e.g., in-swinging, out-swinging or straight), the ball delivery zone (towards the near post, far post, or center of the goal), the number of attacking players involved and whether they were 'static' or 'dynamic' [1,3,4,5,6]. In particular, [1] found that a corner was more likely to be scored on the second ball (after a previous touch from teammate) than directly from the corner kick. From a defensive perspective, they found that hybrid systems – mixtures of man-to-man and zonal marking – concede the most dangerous shots compared to purely zonal or man-to-man marking systems. [4] analyzed the factors that lead to a shot on goal, finding that match time, the number of intervening attackers and whether the attack was 'dynamic' were the most significant variables.

The advent of player tracking data has made it possible to perform a detailed analysis of the synchronized runs made by the attacking players – the rehearsed *routines* that define the core of offensive corner strategy. By identifying and classifying distinct corner routines, we can find those that most frequently create high-quality scoring opportunities. **Using statistical and machine learning techniques, we have developed tools to classify the coordinated runs made by the attacking players during corner kicks, enabling us to identify the distinct corner routines employed by teams in tracking data.**

[1] In the 2019-20 Portuguese Primeira Liga season, just 45 goals were scored from 3082 corners: a 1.5% conversion rate. In the German Bundesliga the rate was 1.9%, and just 0.8% in Spain's La Liga.

Tracking data can also be used to identify and classify the defensive strategies used by teams to repel corner kicks. While teams are generally described as using either a zonal, man-to-man or hybrid system, [1] demonstrated that 80% of the teams in the English Premier League use a hybrid system. However, 'hybrid' encompasses a wide range of defensive options, most notably in the number of zonal defenders and where their zones are located. **To study defensive strategy in more detail, we have developed a supervised classification algorithm to identify the roles of *individual defenders* in corner kick situations.** We use our role classification algorithm to demonstrate that some hybrid systems are more effective than others in repelling corner kicks crossed into the area.

This work has numerous practical applications. Our methodology enables us to reconstruct an opponent's corner kick *playbook* using data from any number of their previous games. Offensively, we identify the distinct routines used by a team, identifying the key runs made by the attacking players and assessing the quality of chances created. Defensively, we can identify which defenders typically mark man-to-man and which mark zonally, study where the zonal defenders are positioned, and assess the strengths and weaknesses of different defensive systems. Finally, a useful by-product of our methods is a new system for encoding attacking player runs; we discuss how this can be used to help players quickly learn new corner kick routines.

2. Data

We make use of tracking and event data for 234 matches from a single season of an elite European professional league. The tracking data for each match consists of the positions of all 22 players and the ball, sampled at a frequency of 25Hz. Individual player identities are tagged in the data, enabling tracking of each player over time. The event data consists of a log of every on-ball action that took place during a match (e.g., passes, shots, tackles and interceptions), the identity of the players involved and the time and location on the pitch at which the event occurred.

We use the event data to provide an initial estimate for the time at which corner kicks occurred. To identify the exact frame at which the corner was taken, we use a combination of factors including the acceleration of the ball and a `ball-in-play` flag included in the tracking data. After removing short corners (which are not included in this analysis) and a small number of corners for which there were ball tracking errors, we were left with a sample of 1723 corner kicks. Finally, to aid comparisons, we reflected the positions of the players and the ball so that corners always appear to be taken from the left side of the field.

3. Classifying corner routines

Figure 1 illustrates two examples of corner routines in our sample. The offensive strategies demonstrated in each example are distinct, particularly in the starting positions of the players, their trajectories and the delivery target of the ball. One of the main objectives of this work is to develop tools to search tracking data to identify the unique corner routines used by a team over many matches. We achieve this by developing a classification system to describe corner routines in terms of the runs made by the players in the attacking team.

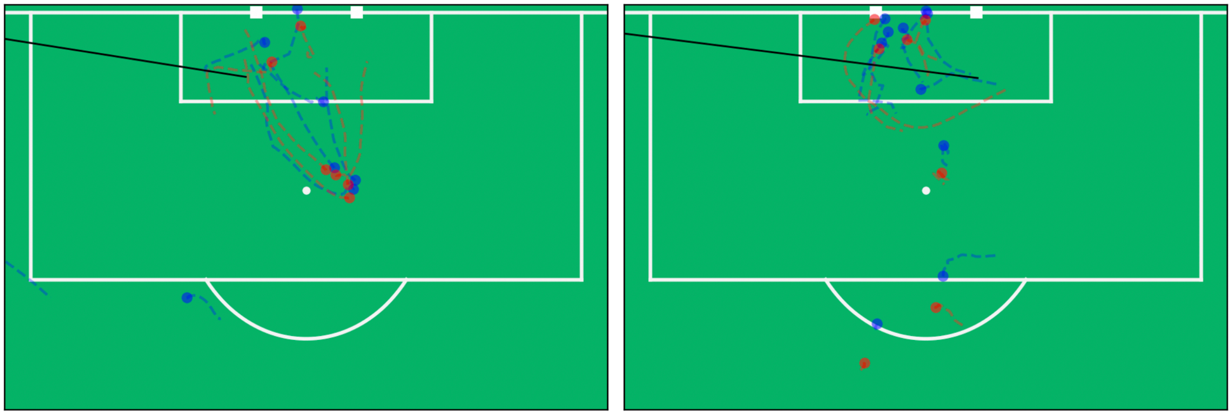

Figure 1: A graphical representation of two different corner routines in our sample. The red (blue) markers indicate the positions of the attacking (defending) players two seconds before the corner is taken. The dashed lines indicate each player's trajectory as the ball is crossed into the area. The solid black line indicates the path of the ball.

Our methodology for analyzing offensive corner routines has two steps:

1. Gaussian mixture modelling to classify attacking player runs into tuples based on their start and end locations; and
2. a topic model (using non-negative matrix factorization) to identify runs that frequently co-occur in corner routines.

We now describe each of these steps in more detail. Note that the trajectory of the ball does not feature in our system. This is because the ball does not always reach the intended target, either because it was intercepted by a defending player or because the cross was not sufficiently accurate.

3.1. Classifying player runs

The basic building blocks of a corner routine are the individual runs made by the attacking players. We define a run entirely in terms of the initial and target location of each player; we do not attempt to model their full trajectory. Initial positions are measured exactly two seconds before the corner is taken, which corresponds to the moment at which the average speed of the attacking players starts to rise as they begin their runs. The target locations are defined as being either the positions of the players exactly one second after the first on-ball event following the corner, or two seconds after the corner is taken, whichever occurs first[2]. It is impossible to know the true target location of a player, we simply assume that attacking players always reach their intended target.

We allocate players to distinct pairs of zones based on their initial and target locations. These zones are defined using the distribution of the initial and target positions of all the attacking players over our entire sample of corners. The process starts with the distribution of the target positions. The upper panel of Figure 2 shows the target positions of nearly 15000 attacking players measured over the 1723 corners in our sample (only players in the attacking quarter of the field are plotted). The large cloud of points in the top-left corner corresponds to the positions of the corner-takers shortly after each corner is taken.

[2] Measuring positions one second after the first on-ball event following the corner kick helps to identify the target position of players aiming to reach the ball following a touch from a teammate.

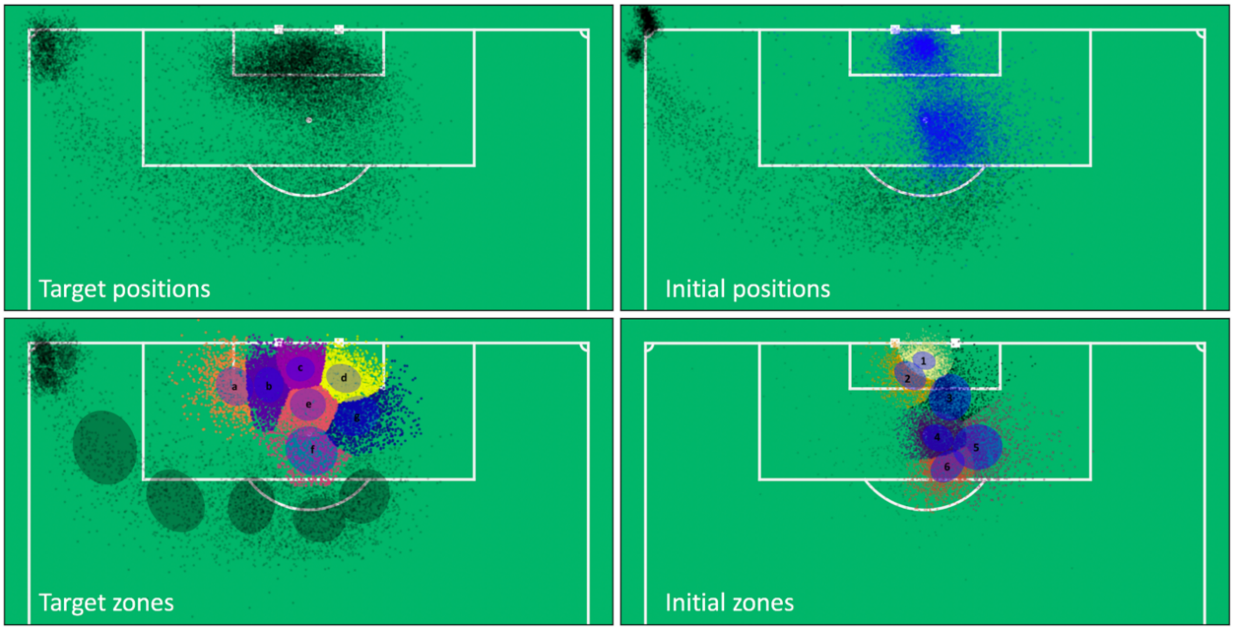

Figure 2: (*upper left plot*) The target positions of all ~15000 attacking players in our sample of corners. (*lower left*) The results of the 15-component GMM fit to the target positions - the seven 'active zones' in the penalty area are represented by blue ellipses and labelled *a-g*. (*upper right*) The initial positions of all 15000 attacking players - players colored blue are tagged as 'active'. (*lower right*) The results of a 6-component GMM fit to the initial positions of the active players.

Target zones are defined by fitting a 15-component Gaussian Mixture Model (GMM) using the expectation-maximization algorithm [7,8]. We find that 15 components (that is, 15 bivariate normal distributions) are sufficient and that adding further components does not result in a significant improvement in the log-likelihood of the fit. The lower-left panel of Figure 2 shows each of the 15 components in the GMM. The seven components of the model located in the penalty area are indicated by blue ellipses and labelled *a* to *g*: we henceforth refer to these as the *active zones*. Individual points belonging to an active zone are colored accordingly. Players with a target position near one of these seven active zones are assumed to be directly involved in the corner routine: these are referred to as *active players*. Players that do not end their runs near an active zone are ignored in the remainder of this work.

The upper-right panel of Figure 2 shows the initial positions of attacking players, two seconds before the corner is taken. *Active players* are colored blue and form two groups: the players starting inside the six-yard box and the players that are initially clustered around the penalty spot. Points colored black are players that were not actively involved in the corner (including the corner-taker, who is no longer involved after taking the corner). To define the initial zones of active players we fit a 6 component GMM model to their initial positions (iteratively removing outliers). The six components of our fit are labelled *1-6* in the lower-right panel of Figure 2.

Allocating players to initial and target zones enables a simple encoding of player runs. Active players are assigned to an initial zone (*1-6*) and a target zone (*a-g*) based on their initial and target positions. For example, in the left panel of Figure 1, the four attacking players that start their runs next to the penalty spot are initially in zone 4, running to target zones *b, c* and *d*. Their runs are therefore encoded as {*4b,4b,4c,4d*}. In total there are 42 possible runs in our system, corresponding to all pairwise combinations of the 6 initial zones and 7 active target zones.

All the runs made by the attacking players during a corner kick can be represented by a 42-element vector in which each element corresponds to a unique run, with its value being the number of players that made that run. Rather than assigning each player to a single run type, we make use of the GMM weights to calculate the probability that they made each of the 42 run types by multiplying the probability that they started in the corresponding initial zone and ended in the corresponding target zone. This process is described in more detail in Appendix 1.

3.2. Topic modelling of run combinations

The runs made by the attacking players are coordinated and synchronized: some players will attempt to draw away defenders, while others will attempt to intercept the ball. The second step of our method is to identify the types of runs that are frequently combined in the same routine. To achieve this, we draw inspiration from topic modelling, making the analogy between runs and words, combinations of runs and topics, and corner kicks and documents.

We use non-negative matrix factorization (NMF) to find a representation of the corners in our data in terms of a basis set of run combinations [9,10,11]. NMF approximates an initial matrix, called the *term matrix* as the product of two lower-rank, non-negative matrices W and H. Our term matrix has the following dimensions: 42 rows by 1723 columns. The rows represent all 42 combinations of the 6 initial and 7 target zones, and each column represents a corner in our data set. W represents the run combinations that frequently co-occur in the data; H tells you how to construct each corner in the data from those run combinations. Further technical details are provided in Appendix 2.

We find that the corners in our data can be accurately reconstructed from a set of 30 run combinations (henceforth referred to as *features*); these features are shown in Figure 3. In some cases, a feature consists of just a single run – this is because the same run may occur in many different types of corner routines.

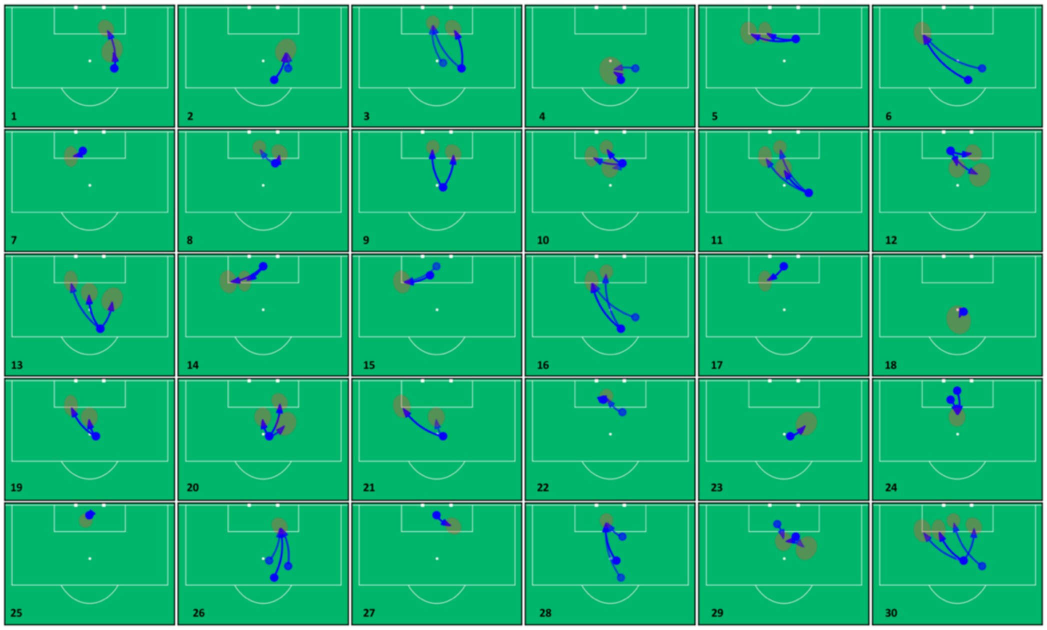

Figure 3: The thirty features, or frequently co-occurring runs, identified by our topic model.

Every corner in our sample can be described in terms of these features. For example, the corner depicted in the left panel of Figure 1 strongly exhibits features 9 and 19, which describe the runs from the penalty spot towards the near post, goal center and far post. The corner depicted in the right panel of Figure 1 strongly exhibits feature 12, which describes the run from the near post towards the far post. Both corners also exhibit feature 25 – an attacking player standing close to where the goalkeeper would be located.

Encoding corners in the feature space enables us to rapidly search very large samples of corners to find those that exhibit a certain feature of interest, or a combination of features. Figure 4 shows other corners in our sample that strongly exhibit feature 12: there are clear similarities between the routines depicted, particularly the curving run made by a player from one goal post towards the other. One team in particular in our dataset made frequent use of the routines shown in Figure 4.

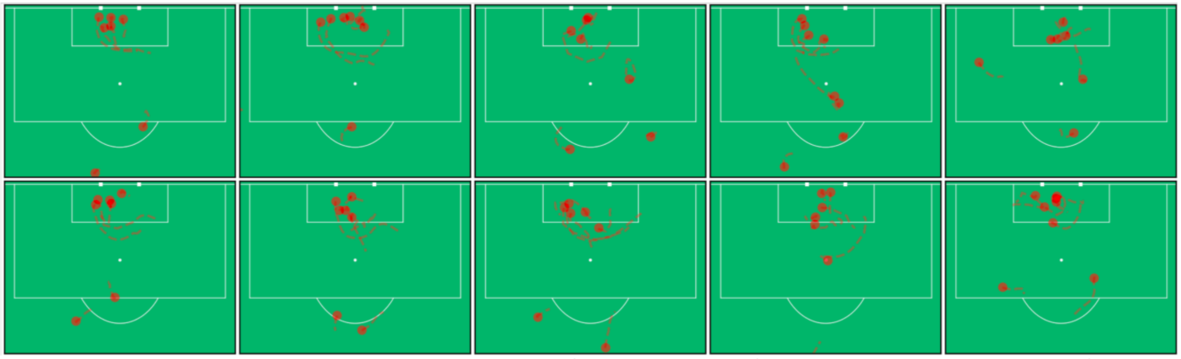

Figure 4: Ten corners in our sample that strongly exhibit feature 12 – runs from the near post round to the far post. Only attacking team players are shown. Dots indicate the initial positions of each player and dashed lines show their trajectories.

Distinct corner routines in our dataset can be identified by grouping corners that exhibit similar feature expressions (the columns of the H matrix) using agglomerative hierarchical clustering. We present the distinct corner routines found for individual teams in our data in Section 5.1.

4. Identifying defensive roles

The defining feature of defensive strategy during corners is the use of man-to-man and zonal marking. Man-to-man marking requires a player to closely track a specific opponent, while zonal marking requires a player to defend a spatial region. Teams are frequently described as adopting either man-to-man, zonal, or 'hybrid' (mixed) systems to defend corners. However, few teams use an exclusively zonal system, and it is rare for a team to have no zonally marking players whatsoever[3]; most teams use a hybrid system to defend corner kicks [1]. Video examples of hybrid systems used in the 2019-20 UEFA tournaments can be found here, here and here.

In this work, we do not seek to classify the defensive system of teams as a whole. Instead, we work on the level of individual players, using supervised machine learning to classify the role of *each defender*. This provides significantly more information about the nature of a defensive system. We use the *XGBoost* [12] implementation of gradient-boosted decision trees to calculate the probability

[3] The number of defenders typically exceeds the number of active attacking players. so at least 1 defender is free to zonally mark.

that each defender in the penalty area is marking man-to-man (vs zonally). Gradient boosted decisions trees have been demonstrated to be a powerful tool for solving classification prediction problems with many predictors in an efficient manner [13].

Players marking man-to-man will typically start in close proximity to an opponent and will often cover a significant distance during the corner kick as they track the opposing player. Zonally marking players, on the other hand, tend to be more stationary. As Figure 2 demonstrates, the initial positions of attacking players in the penalty area are grouped into two clusters: we therefore expect man-to-man marking defenders to be initially positioned in (or, at least, near to) one these clusters. Working with SL Benfica's analysts, we selected a set of ten metrics to predict the roles of each defender[4]; these metrics are described in Appendix 3.

To provide data for training and testing, analysts at SL Benfica watched 500 movies of corners randomly selected from our data set and manually identified the jersey numbers of the man-to-man and zonally marking defenders. The resulting data set consists of 3907 defenders: 55% were tagged as marking man-to-man and the remainder tagged as marking zonally; the classes are therefore well-balanced. Figure 5 shows the initial positions of zonal (upper-left panel) and man-to-man marking players (upper-right) in the analyst's sample. As anticipated, a defender's initial position is a strong indicator of their defensive role. Note that the distribution of man-to-man marking defenders resembles the spatial distribution of attacking players shown in the top-right panel of Figure 2.

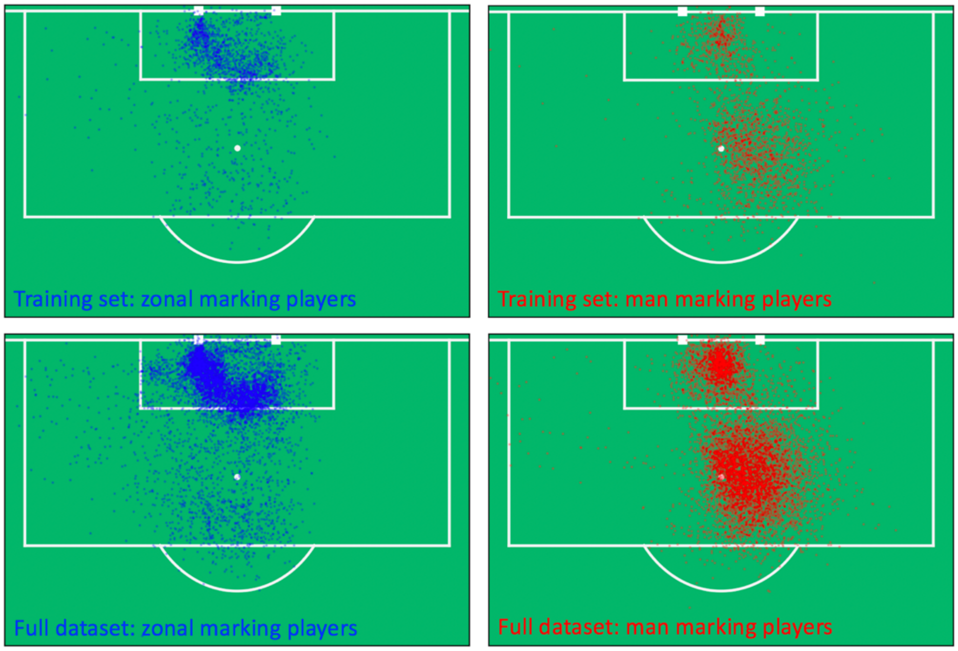

Figure 5: *(upper row)* The initial positions of the zonal (*left*) and man-marking players (*right*) in our training sample, based on the manual classifications made by SL Benfica's analysts. (*lower row*) Results for the full sample using the classifications predicted by our boosted decision tree.

[4] Note that our current methodology does not explicitly identify which attacking player a man-to-man marking defender is marking.

Training XGBoost with 10-fold cross validation resulted in a classification accuracy of 83.4% ± 2.1%[5]. Therefore, the roles of approximately five in every six defenders in our training sample were correctly classified. The most predictive metrics, selected via their F-score, are discussed in Appendix 3. We applied the decision tree to our remaining sample of 1223 corners to predict the roles of every defending player. The initial positions of players classified as zonally marking (lower-left) and man-to-man marking (lower-right) are shown in Figure 5. There is clearly a strong resemblance to the analyst-annotated sample. Two video examples of defensive role classifications can be found here.

Once we have identified which defenders were allocated to zonal marking roles, we can investigate where their zones were located. Figure 6 shows a heatmap of the spatial distribution of the zonal defenders for Team Y over the 130 corner kicks they faced during the 2017-18 season. A darker shade of blue indicates a higher occupation rate for a given position.

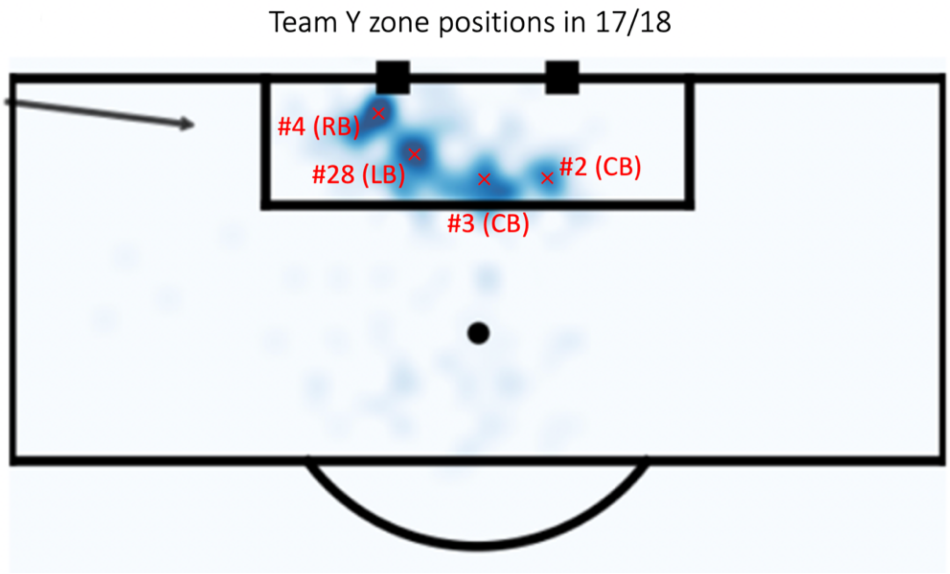

Figure 6: The spatial distribution of zonally marking defenders for Team Y during their matches in the 2017-18 season. The arrow indicates the direction from which the corner is taken.

Team Y used a defensive system consisting of four zonally marking players and four players marking man-to-man. The heatmap clearly indicates the locations of the four zones: there are four distinct peaks in the distribution, shielding the goalmouth from the ball. The red crosses indicate the most likely positions of each defender. The identities of the four zonal defenders remained largely the same throughout the season: the right-back (#4) was positioned at the near-post, the left-back (#28) stationed next to him, with the two center-backs (#3 and #2) defending the edge of the six-yard box. We discuss how this information can be used to assess the strengths and weaknesses of a zonal system in the following section.

[5] 81% of the players classified as zonally marking by the model were tagged as such in the training set (precision) and 81% of the players tagged as zonal markers in the training set were correctly identified (recall).

5. Practical Applications

5.1 Analysis of an opponent's offensive corner strategies

Anticipating how an opponent will play in different situations is a crucial component of pre-match preparation. A well-prepared analysis can significantly increase the chance of a positive outcome [14,15]. However, evaluating an opponent using video is a time-consuming process and so club analysts are often limited to watching up to five matches to inform their reports for the coaching staff. Such a small sample of matches provides only a limited insight into the range of set-piece strategies utilized by an opponent.

Our methodology enables us to rapidly identify the key features of the corner strategies used by teams, both offensively and defensively, over hundreds of matches. In this section we demonstrate its application to opposition analysis, highlighting the distinct corner routines used frequently by four teams over the course of a season. Identifying specific examples of particular corner routines enable analysts to create a video montage for coaches. Our tools are already being employed by two professional teams (SL Benfica in the Portuguese Primeira Liga and Aalborg FC in the Danish Superliga) to analyze their opponents in upcoming games.

Figure 7 shows examples of the corner routines used regularly by teams A (red), B (blue), C (green) and D (black) in the 2017/18 season. Each panel shows a specific example of a distinct routine that the team used multiple times throughout the course of the season. The circles indicate the starting position of each player and the dashed lines indicate their runs in the seconds that follow the corner kick. The black solid line indicates the trajectory of the ball.

A popular strategy used by almost every team in our sample is the *jellyfish*. In this strategy, three or four players start in a cluster outside the six-yard box before making gradually diverging runs towards the box (see the plots in rows 3 and 4 in Figure 7). Closer inspection reveals that the teams used different implementations of this strategy, varying the position of the initial cluster and the length of the runs made by each player. One example is the *train*, in which players start in a line rather than a cluster, as popularized by the England team at the 2018 World Cup. Team B regularly employed the *train*, with the line starting near the penalty spot (third row); they also used a variation of the *jellyfish* in which the cluster of players was positioned in the far corner of the penalty area (fourth row). Team A employed a variation in which a player made a run around and behind the initial cluster, aiming for the far post (fourth row) and an unusual routine in which four players start at the far edge of the six-yard box before running horizontally towards the ball (second row).

Another class of routines is the *overload*, in which four or five attacking players are initially positioned to crowd the six-yard box very close to goal. The second row of Team D shows an example in which two players positioned in front of the near post make runs out of the six-yard box and round towards the far post to intercept a deep delivery. The first row for Team D shows the reverse of this: two players at the far post run around the box to intercept a near post-delivery. Team B also regularly employed a variant of the *overload* (first row).

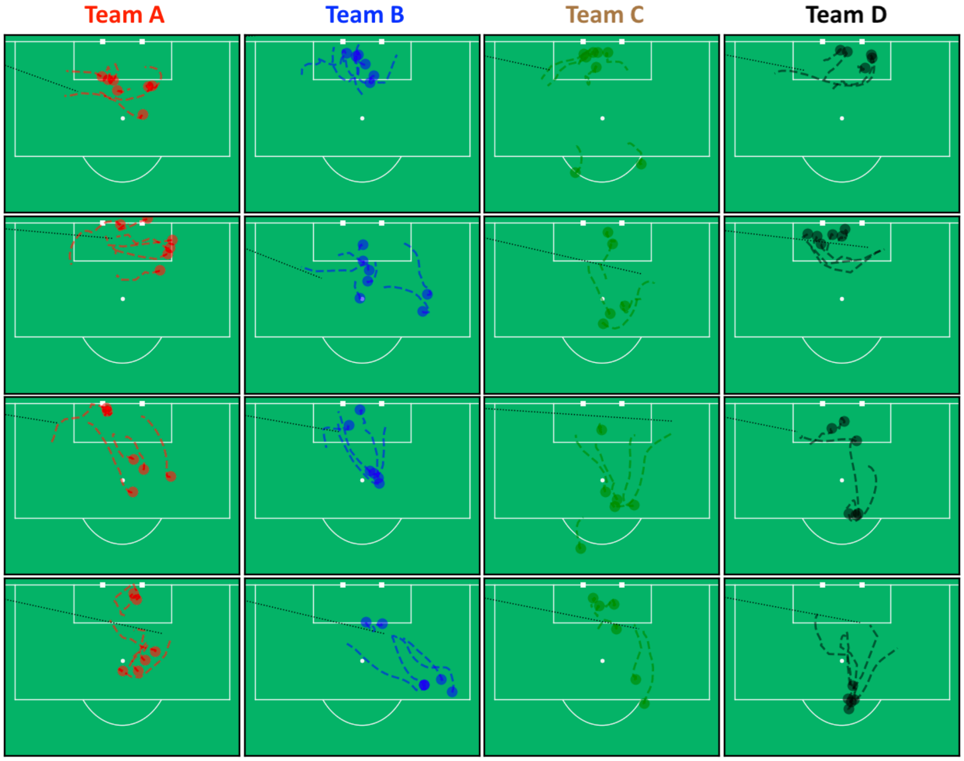

Figure 7: Examples of popular corner routines employed by four teams in our dataset.

The teams in our data did not alternate randomly from one routine to another as they took corners. Rather, they would use a routine regularly over a consecutive series of matches and then discard it, perhaps reintroducing the routine later in the season. For example, Team C attempted the corner routine depicted in the fourth row five times over three consecutive games, discarded it for six games and then used it three times in one game. This emphasizes the need to scan over a large number of matches to fully scout the range of corner strategies that might be employed by an opponent in their next match.

5.2 Comparing the effectiveness of zonal systems

Most teams choose to defend corners using a mixture of zonal and man-to-man marking players; however, coaches must still decide on the number of zonal defenders to use and where their zones should be located. When making these decisions, a coach will naturally want to know how frequently (and where) a particular combination of zones concedes shots and goals. By automatically identifying zonally marking defenders in tracking data, our methodology enables us to analyze the *effectiveness* of different configurations of zonal defenders. We now provide an example.

Figure 8 compares the shots and goals conceded by two different zonal systems, one with four zonal defenders (left panels) and one with two zonal defenders (right panels), extracted from a sample of 3500 corners encompassing two seasons of tracking data. The four-zone system was used in 366 corner kicks by eight different teams, and the two-zone system in 600 corners by ten different teams. Both systems also included man-to-man marking defenders (not shown).

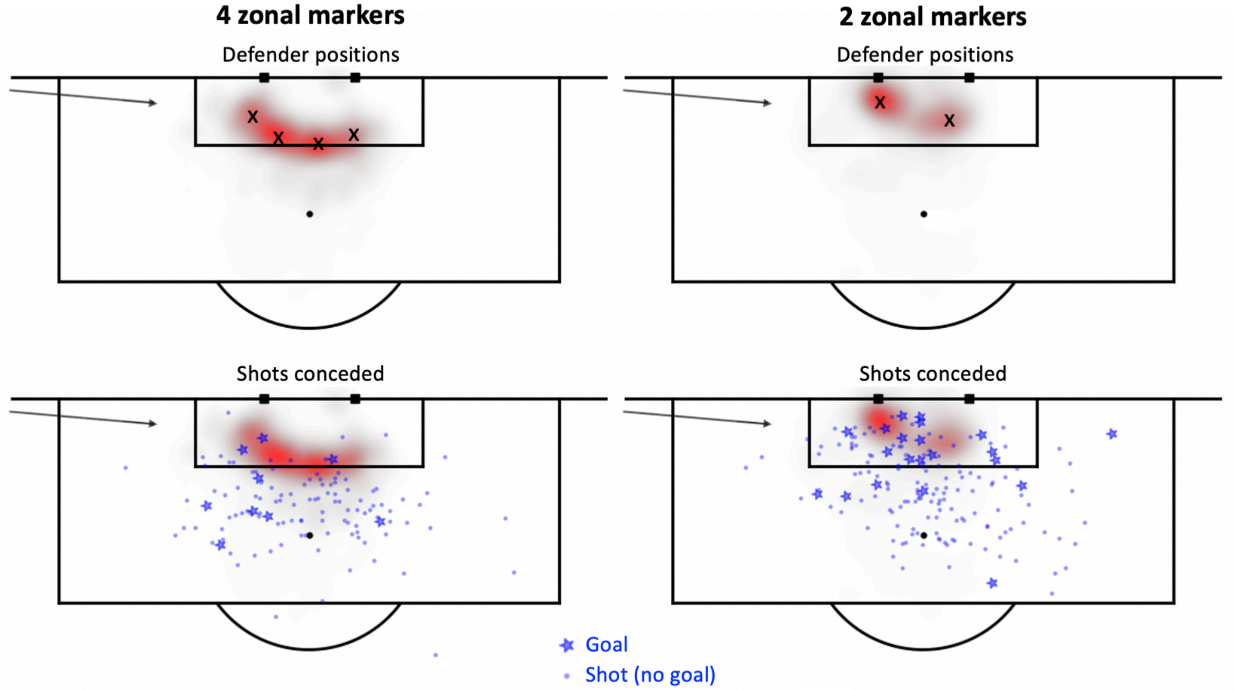

Figure 8: Comparison of the shots conceded by two zonal marking systems: 4 zonal markers (*left plots*) and 2 zonal markers (*right*). The upper plots show the distribution of the positions of the zonal defenders in each system; the lower plots indicate the delivery location of corners in which shots or goals were conceded.

The top row of Figure 8 shows the positioning of the zonal defenders in each system. The intensity of the color is proportional to the probability that a defender was occupying that position at the instant a corner was taken. In the four-zone system (top-left) the four defenders are positioned in a ring around the goal mouth: the x's indicate their most likely positions. In the two-zone system, the zones are located at the near post and slightly beyond the center of the 6-yard box. Note also that the zones in the two-zone system are positioned closer to goal than in the four-zone system.

The lower panels in each plot show the delivery location of corners that resulted in a shot on goal within 6 seconds of the corner being taken (stars indicate goals, dots indicate shots that missed the target or were saved). The four-zone system conceded shots in 33% of the corners that it faced, while the two-zone system conceded shots in 27% of corners faced. However, the *quality* of the chances conceded were significantly lower in the four-zone system: 7.5% of the shots conceded produced a goal, compared to 13% in the two-zone system. The reason for this is clear: in the four-zone system, most of the shots allowed came from corners delivered outside of the 6-yard box (lower-left panel of Figure 8). Very few shots occurred from corners delivered into the region between the zonal defenders and the goal line. In the two-zone system, one-third of the shots allowed (and two-thirds of goals conceded) were from corners delivered into the 6-yard box. Closer inspection of the lower-right panel of Figure 8 shows that many of the goals conceded were from corners delivered either directly between the two zonal players or behind the second zonal player.

The analysis in this section demonstrates that, while the four-zone system conceded slightly more shots on goal, those shots tended to be from a greater distance than the shots conceded by the two-zone system and were therefore less threatening. Overall, the two-zone system conceded goals at a higher rate than the four-zone system and so can be considered the less effective of the two systems. This analysis can be extended to any zonal system used regularly by professional teams.

5.3 Training optimization

Teams playing in both domestic and continental competitions must often deal with congested schedules, sometimes playing matches at a rate of 3 every 9 days. During this period, the days between matches are reserved for recovery and for giving the players who didn't play sufficient physical load. Time working on strategy in training is a limited resource and many teams dedicate less than an hour to practicing set pieces. An important application of our tools is to increase the impact of the time that is committed to practicing set pieces in training. This can be achieved in two ways.

First, as demonstrated in the previous section, our data-driven approach enables rapid identification of the strengths and weaknesses of an opponent's defensive system at corner kicks. Time spent practicing corner kicks in training can then focus specifically on the corner routines that are mostly likely to enable a player to evade the defence and take an unchallenged shot at goal. Additionally, coaches can brief players on the different corner routines used regularly by their next opponent, helping the players to identify a routine from the opponent's starting positions and anticipate the likely delivery target of the ball.

Second, our system for encoding player runs based on specific *initial* and *final* zones introduces an intuitive method for helping players to learn new corner routines. Each player needs to remember only the alphanumeric code for their own run (e.g., *4b, 2a*, etc, as described in Section 3.1) and the delivery area of the ball for each new routine. This allows coaches to efficiently introduce new corner routines that won't have been seen by an opponent's video analysts.

Nuno Mauricio, the head of Match Analysis for SL Benfica, states: "*Corner kicks are perhaps a moment of the game that is dominated by strategy, rather than a balance of creativity, tactical behavior and strategy as in open play situations. This sort of analysis helps a coach to understand which kind of corners are more effective against certain defensive setups, thus having a direct impact on the training process and the development of the sport itself*". The implication is that the rewards of practicing corner kicks are directly proportional to the research and planning that goes into them, and less dependent on the instinctive, split-second creativity of players during a match.

6. Summary and future work

Using player tracking and event data, we have conducted an in-depth analysis of the offensive and defensive strategies employed by teams in corner kick situations. By studying and classifying the runs made by the attacking team's players, we have created a 'language' for corner kicks that enables us to characterize each routine in terms of a distinct set of run combinations. This allows us to rapidly search a large sample of corners for certain characteristics (such as a particular run, or combinations of runs) or to find the distinct corner routines used by a particular team.

We have also presented a supervised learning model for classifying the role of each defending player in corner situations. Using a sample of 500 corners manually annotated by SL Benfica's analysts, we trained the XGBoost algorithm to predict whether each defending player was instructed to mark man-to-man or zonally, obtaining a cross-validated classification accuracy of 83.4% ± 2.1%.

We have demonstrated how these tools can be applied to provide unprecedented insights into the strategies used by teams in corner kick situations, identifying the distinct corner routines employed by individual teams over the course of a season and quantifying the strengths and weaknesses of different defensive systems.

A natural next question to ask is: *which attacking routines are most effective against a certain defensive set-up*? We have refrained from providing an empirical answer to that question in this paper because of the limited size and scope of our data set. In a follow-up work we will make use of a significantly larger sample of data to empirically investigate the most effective strategies for increasing the quality and quantity of chances created in corner kick situations.

7. Acknowledgements

We acknowledge Devin Pleuler at Toronto FC for his advice and insights, and Tiago Maia and Jan Schimpchen from SL Benfica for helping to produce the training data for our defensive role classification model.

References

[1] Power, P., Hobbs, J., Ruiz, H., Wei, X., Lucey, P. *Mythbusting Set-Pieces in Soccer* In Sloan Sports Analytics Conference, 2018

[2] The Athletic, *Set-piece kings Midtjylland – and their former Celtic star – have left top teams admiring from afar*, Available: https://theathletic.com/1103112/2019/08/07/how-set-piece-kings-midtjylland-and-their-former-celtic-star-have-left-even-the-likes-of-manchester-city-behind/

[3] Hannah Beare & Joseph Antony Stone. *Analysis of attacking corner kick strategies in the FA women's super league 2017/2018* In International Journal of Performance Analysis in Sport, 19:6, 893-903, 2019

[4] Casal, C. A., Maneiro, R., Arda, T., Losada, J. L., & Rial, A. *Analysis of corner kick success in elite soccer* In International Journal of Performance Analysis in Sport, 15, 430–451, 2015

[5] Pulling, C. *Long corner kicks in the English premier league: Deliveries into the goal area and critical area* In Journal of Fundamental and Applied Kinesiology, 47(2), 193–201, 2015

[6] Pulling, C. & Newton, J. *Defending corner kicks in the English Premier League: Near-post guard systems* In International Journal of Performance Analysis in Sport, 17(3), 283–292, 2017

[7] A. P. Dempster, N. M. Laird, and D. B. Rubin, "*Maximum likelihood from incomplete data via the EM algorithm*", Journal of the Royal Statistical Society: Series B(Methodological), vol. 39, no. 1, pp. 1–22, 1977.

[8] Friedman, J., Hastie, T., & Tibshirani, R. *The elements of statistical learning,* Vol. 1, No. 10. New York: Springer series in statistics, 2001

[9] da Kuang, D., Choo, J., & Park, H. *Nonnegative matrix factorization for interactive topic modelling and document clustering*. Partitional Clustering Algorithms (pp. 215-243). Springer International Publishing, 2015

[10] D. D. Lee & H. S. Seung. *Algorithms for non-negative matrix factorization*. Advances in Neural Information Processing 13 (Proc. NIPS*2000) MIT Press, 200110

[11] P. O. Hoyer. *Non-negative matrix factorization with sparseness constraints*, Journal of Machine Learning Research,5:1457–1469, 2004

[12] Chen, T. & Guestrin, C. *Xgboost: A scalable tree boosting system* In Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 785–794, 2016

[13] Sagi O, Rokach L. *Ensemble learning: A survey*. WIREs Data Mining Knowl Discov., 2018

[14] Carling, C., Williams, M., & Reilly, T. *Handbook of soccer match analysis: A systematic approach to improving performance* In London: Routledge, 2005

[15] Sarmento, Hugo Marcelino, Rui Campanico, Jorge Matos, Nuno & Leitão, José. *Match analysis in football: a systematic review.* Journal of sports sciences. 32. 1831-1843, 2014

[16] Févotte, C., & Idier, J. (2011). *Algorithms for nonnegative matrix factorization with the β-divergence.* Neural computation, 23(9), 2421-2456.

Appendix 1 – Encoding corners

The runs made during a corner kick can be represented by a 42-element vector, x_a, in which each element corresponds to a unique run (combination of initial and target zones). The value of x_a for any $a \in [1,42]$ is determined by calculating the product of the probability of each player occupying the corresponding initial and target zones and then summing up the probability over all active attacking players. The formulation is as follows:

$$x_a = \sum_{p=1}^{N_p} \mathrm{P}(\mathrm{R}_a) = \sum_{p=1}^{N_p} \mathrm{P}_p\left(InitialZone_{\mathrm{R}_a}\right)\mathrm{P}_p\left(TargetZone_{\mathrm{R}_a}\right) \tag{A1}$$

where $N_p$is the number of active attacking players, $\mathrm{P}_p\left(InitialZone_{R_a}\right)$ is the probability (i.e. the GMM mixture weight) that player p started in the initial zone of run R_a and $\mathrm{P}_p(TargetZone_{R_a})$ is the probability that player p ended in the target zone of run R_a.

This probabilistic method for encoding runs therefore allows for uncertainty in the precise run type that a player made.

Appendix 2 – Non-negative matrix factorization

Non-negative Matrix Factorization is a linear algebra method that decomposes a high dimensional non-negative matrix into two lower rank factor matrices. Let X be a n x p non-negative matrix where element $x_{ab} \geq 0 \ \forall \ x_{ab} \in X$. The rows of X in our case correspond to the 6 x 7 = 42 combinations of initial and target zones, while the columns represent each of the 1723 corners in our data set (i.e., each column is the encoding of a unique corner in the data).

Non-negative Matrix Factorization aims to find an approximation

$$X \approx WH \tag{A2}$$

where W and H are n x r and r x p non-negative matrices, respectively. The factorization rank r is chosen such that $r < min(n,p)$ and represents the size of the new basis set (the number of distinct run combinations, or features, in our corner sample). Henceforth, we will refer to W as the basis matrix and H as the coefficient matrix.

We solve for W and H by minimizing a distance function:

$$\frac{1}{2}\|X - WH\|_F^2 + \ \alpha||W||_F^2 + \alpha||H||_F^2 \ , \tag{A3}$$

the Frobenius norm of the difference between X and the product of the W & H matrices, plus the sum of the norms of the W and H matrices, each multiplied by a regularization parameter α, which we set to 1. The regularisation encourages sparseness in the basis set vectors, which is helpful because the number of attacking players actively involved in corners is normally in a fairly narrow range (5-8). The distance function is minimized using a multiplicative update procedure, where W and H are iteratively and alternately updated in a methodology not dissimilar to Expectation-Maximisation [10,16].

We chose the number of features (the value of r) by inspecting how $||X - WH||_F^2$ decreases as the size of the basis set, r, is increased from 1 to 42. We selected r = 30, which is where the gradient of this curve approaches zero.

Appendix 3 – XGBoost Features

The key distinction between man-to-man and zonal marking is that, in the former, a player is marking a moving target rather than a static region. Treating attackers near the goalkeeper as a special case and emphasizing the locomotive reaction of the defenders, we selected the metrics listed below as predictive variables for our XGBoost classifier [12]. These variables were vetted by video analysts at SL Benfica (note that our methodology does not identify the specific opponent a defender is man-marking). The most predictive features, selected via their F-score, are marked with an asterisk in the list above:

1. initial position (x coordinate) *
2. initial position (y coordinate) *
3. distance between start and target positions *
4. initial proximity to goalkeeper *
5. average distance travelled by attacking players in the same initial zone *
6. average distance travelled by other defenders in the same initial zone *
7. initial zone
8. target zone
9. number of attacking players in the same initial zone
10. number of other defenders in the same initial zone

There are several variables that are important in predicting role classifications but, as Figure 5 demonstrates, even just the initial positions of the defenders are a useful discriminator. Using only the initial positions as independent variables in a simple logistic regression classifier provides 70% classification accuracy for the defensive roles in our training set. Adding the distance travelled between player start and target locations as a variable in the logistic regression increases the classification accuracy to 74%.