Towards Expected Counter - Using Comprehensible Features to Predict Counterattacks

Henrik Rolf Biermann[0000−0001−5660−9876] [1], Franz-Georg Wieland[2,3] [0000−0002−5629−8622], Jens Timmer[2,3,5] [0000−0003−4517−1383], Daniel Memmert[1] [0000−0002−3406−9175], and Ashwin Phatak[1] [0000−0003−1772−8949]

[1] Institute of Exercise Training and Sport Informatics, German Sport University of Cologne, Cologne 50933, Germany
h.biermann@dshs-koeln.de
[2] Institute of Physics, University of Freiburg, Freiburg, Germany
[3] Freiburg Center for Data Analysis and Modelling (FDM), University of Freiburg, Freiburg, Germany
[4] Germany Centre for Integrative Biological Signalling Studies (CIBSS), University of Freiburg, Freiburg, Germany

Abstract. Soccer is a low-scoring game where one goal can make the difference. Thus, counterattacks have been recognized by modern strategy as an effective way to create scoring opportunities from a position of stable defense. This coincidentally requires teams on offense to be mindful of taking risks, i.e. losing the ball. To assess these risks, it is crucial to understand the involved mechanisms that turn ball losses into counterattacks. However, while the soccer analytics community has made progress predicting outcomes of single actions (shots or passes) [1, 2] up to entire matches [15], individual sequences like counterattacks have not been predicted with comparable success. In this paper, we give reasons for this and create a framework that allows understanding complex sequences through comprehensible features. We apply this framework to predict counterattacks before they happen. Therefore, we find turnovers in soccer matches and create transparent counterattack labels from spatiotemporal data. Subsequently, we construct comprehensible features from sport-specific assumptions and assess their influence on counterattacks. Finally, we use these features to create a simple binary logistic regression model that predicts counterattacks. Our results show that players behind the ball are the most important predictive factors. We find that if a team loses the ball in the center *and* more than two players are not behind the ball, they concede a counterattack in almost 30% of cases. This stresses the importance to avoid ball losses in build-up play. In the future, we plan to extend this approach to generate more differentiated insights.

Keywords: Soccer Analytics · Spatiotemporal Data · Event Data · Interpretable features .

1 Introduction

As part of a universal language to describe soccer, the so-called moments of play have been introduced. This concept presents a way to organize the sometimes chaotic developments during matches into five phases: The two attacking phases for each team (where the other team defends), two transition phases between these attacking phases, and set pieces [8].

In performance analysis the transition phases have recently gained attention [13]. Counterattacks, i.e. fast attacks after winning the ball from the opposition have been determined as an important factor for goal scoring and chance creation [10, 14]. Thus, it is essential for a team's success to investigate potentially dangerous situations for counterattacks. However, existing work rather deals with single actions like passes [1, 16] or shots [2] or long-term prediction, i.e. match [15] or competition [7] outcomes. In the mid-term, predictive studies are rare and usually use large feature sets and machine learning which makes it difficult to translate the results into sport-specific insights. Thus, there is a lack of interpretable models that can help to understand the mechanisms involved.

Thus, this study aims to predict counterattacks through comprehensible features. To achieve this, the study proposes a general framework for analyzing complex sequences by building comprehensible features (Section 2). It also describes methodological steps by finding turnovers and defining counterattacks (Section 3). In Section 4 we build comprehensible features and assess their importance for counterattack success, before combining these features into a predictive model in Section 5. Section 6 concludes with a summary of our findings and an outlook on future research topics.

2 Framework for Understanding Complex Sequences

Soccer matches are difficult to predict due to the high number of degrees of freedom and lack of structure compared to other sports [11]. Prediction approaches are either done on very small or very large time scales. Short-term approaches aim to predict the outcome of a single event, e.g. expected goals model the success (goal) probability of a given shot [3] and expected passes model the completion probability of a pass [1]. The advantage of such approaches is that it is performed in an enclosed environment (e.g. a particular shot) with a small number of possible outcomes (e.g. goal yes/no). Long-term approaches forecast the outcome of a match [15] or an entire season/tournament [7]. These techniques benefit from compounding the aggregated complexity of individual processes into a small number of outcomes e.g. of the final score or outcome of a match.

Analysis on an intermediate time scale, however, is less frequently approached. This is mainly due to three reasons. (i) Defining clearly defined sequences is difficult. While many colloquial terms like attacking sequence, pressing moment, and counterattack exist, their translation into precise, rule-based definitions is non-trivial. (ii) Even when these situations are characterized, their meaning, importance, and impact in the broader context of the game are not quantified.

Thus, defining *success criteria* is difficult as objectives vary, e.g. dependent on playing style or game situation. (iii) Lastly, even if both the sequence and the success criteria are clearly defined, individual sequences comprise a variety of individual (concrete and possible) inherently complex actions.

Fernandez et al. [6] have tackled this problem by developing a framework that estimates the value of a possession at any time using conditional probability and the expected outcome of different possible actions. Their general model can be used to evaluate actions during a sequence, however, it requires exact models for individual actions which may vary between game situations. Bauer et al. [2] have identified pressing situations after turnovers using an XGBoost model and a broad range of hand-crafted features. Their resulting model is highly predictive, however, while the presented relative importance of features offers some insights, this information is not easily translatable to sport-specific terms or guidelines.

Thus, we propose a novel framework to approach the prediction of specific sequences based on comprehensible features. This enables objective and transparent analysis and allows to create concrete questions and actual guidelines.

1. Define a precise, rule-based specification of *sequences of interest*.
2. Define precise, data-driven and sequence-specific *success criteria*.
3. Construct *comprehensible features* based on sequence-specific assumptions.
4. Assess the *prediction capability* of these features for the success of sequences.
5. Build a *predictive model* based on previously identified relevant features.

It is important to note, that the comprehensible features constructed in step three of the process are in general not unique in their ability to predict the sequence success. The complexity of football makes it highly likely, that many different features can be constructed with similar predictive capabilities. This holds especially true, since for any non-trivially constructed feature, many slightly altered features can be found. The choice of feature is thus an integral part of the analysis pipeline, with which the hypothesis space to be examined is already circumscribed. Thus, the feature construction itself already allows for a precise specification of different hypothesis, which also makes it possible to assess the importance of certain feature characteristics by using slightly altered feature subsets in the analysis.

3 Definition of Sequences of Interest and Success Criteria

In this section we introduce rules for identifying *sequences of interest* from turnovers (Section 3.1), define a set of sequence-specific *success criteria* for counterattacks (Section 3.2), and apply both to create our dataset from a provided number of matches (Section 3.3).

3.1 Rule-based identification of persistent open-play turnovers

Our provided match data (see Section 3.3) contains position and event data of matches, however, lacks information about possession. Thus, we define a turnover

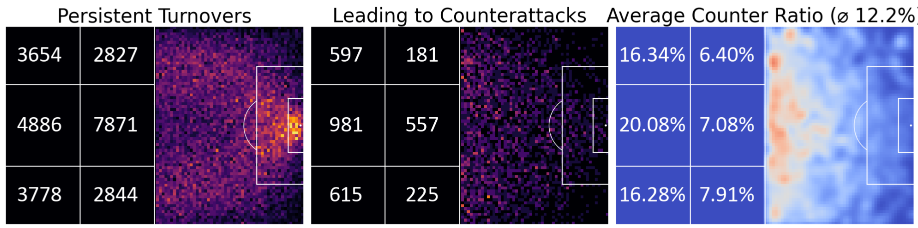

Fig. 1. Influence of location for counterattack probability. Left: Location of *Persistent Turnovers* on the pitch. Middle: Location of *Persistent Turnovers* that turn into counterattacks. Right: Smoothed ratio of counterattacks to total turnovers at every location. For the two left plots the brighter the area the higher the number of occurrences. For the right plot, blue corresponds to low ratios (close to zero) and red regions correspond to high ratios. Dark red corresponds to high ratios up to the maximum value of 40%. Additionally, we divide the pitch into six zones and display the number of turnovers and ratio within a zone.

as a pair of adjacent events that are assigned to players from opposing teams and refer to the latter as the *ball loss* event. This intuitive definition, however, leads to a large number of turnovers that are not all equally relevant to the present study. Turnovers in the own half of the ball losing team are a special case as they generally involve an immediate attack [2]. Turnovers after set pieces (corner, freekick, throw-in, goalkick, etc.) occur in situations of special (unusual) positioning scenarios [8]. Turnovers that happen during a dead ball (ball-out-of-bounds, foul, offside, etc.) present a special case due to the involved stoppage of play. Finally, turnovers where the ball winning team immediately loses control of the ball (e.g. clearance or a subsequent turnover) indicate that there was no coordinated attack performed. Thus, we exclude the following four subgroups of turnovers:

1. *Own Half Turnover*: The ball at the ball loss event is within the own half of the ball losing team.
2. *Dead Ball Turnover*: Ball loss event is a dead ball.
3. *Set Piece Originated Turnover*: Ball loss event is a set piece or there is a set piece within the last $t_1 = 5\text{s}$ before the turnover.
4. *Non-persistent Turnover*: After the turnover, there is either another turnover or a dead ball within the next $t_2 = 3\text{s}$.

We choose the values for t_1 and t_2 based on discussions with domain experts after watching video material. We exclude the four subsets and refer to the remaining as persistent open-play turnovers in the opposition half, or shorter, *persistent turnovers*. These define the start points of our *sequences of interest*.

3.2 Definition of success criteria for counterattacks

A counterattack has been identified by Lago-Ballesteros et al. [10] as a fast and direct attack (with few players) that starts after winning the ball and by

StatsPerform [12] as a team gaining possession and moving the ball into a target area within the opposition's half (where the speed determines its value). A similarity of both of these definitions is that they include a temporal and a spatial component. Thereupon, we derive the following descriptive *success criteria* for counterattacks:

1. *Spatial*: The distance between the ball and the goal of the ball losing team is reduced to less than $d = 35$m.
2. *Temporal*: The spatial criterion is met within a time window of $t_3 = 15$s after the turnover.

However, these definitions do not include possession of a team. Therefore, an uncontrolled clearance (without intent) towards the opponent's half would also be a counterattack. Concerning this, we introduce the additional criterion:

3. *Sustain Possession*: After the spatial criterion is met there is another offensive event (shot, pass, or ball carry) of the ball winning team with a distance of less than $d = 35$m within the next $t_4 = 5$s.

The values for d, t_3, and t_4 are carefully chosen after discussion.

3.3 Emerging Dataset

The provided data by Stats Perform contains 289 matches of position data captured at 25Hz and manually captured event data from a recent season of topflight soccer. We exclude all situations where players are missing (e.g. due to a red card). In the remaining data we find 72,367 turnovers (250 per game) of which 18,364 (64 per game, 25.6%) are dead ball turnovers, 8,195 (28, 11.2%) are set piece originated turnovers, 19,948 (69, 27.6%) are non-persistent turnovers and 25,860 (89, 35.6%) are *persistent turnovers*. In the set of *persistent turnovers*, we find 3,156 (11 per game, 12.2%) turnovers that turn into a counterattack. We divide the pitch into a grid of $1m^2$ squares and use the location of turnovers and counterattack labels to compute the ratio of counterattacks to total turnovers at each location. We perform a smoothing of the ratio according to [9] with smoothing factor of $s = 2.5$. The number of turnovers and counterattacks in the grid and the ratio of counterattacks are visualized in Figure 1.

4 Comprehensible Features for Prediction

In this section we make domain-specific assumptions about favorable scenarios during counterattacks to construct three distinct *comprehensible features* (Section 4.1). Subsequently, we reflect on the influence of the ball loss location (Section 4.2). Finally, we assess each feature's capability to predict counterattacks by comparing counterattack ratios for specific values with the average ratio (Sections 4.3).

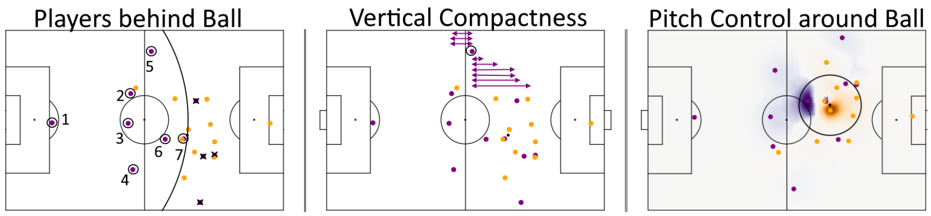

Fig. 2. Computation of the three constructed interpretable features for an exemplary turnover situation. Displayed are the positions of the ball (black), the ball losing team (purple), and the ball winning team (orange). Left: Losing players behind ball are circled. Middle: The vertical distance of the circled player to all teammates is indicated. Right: Pitch control surface [5] with exponentially decreasing weights around the ball is shown and the most relevant area for computing pitch control is circled.

4.1 Constructing Features from Domain-Specific Assumptions

As discussed in Section 2 the features constructed below are not unique and slight variation of these features can as well be used with comparable results. Therefore, we try to report our specific domain-specific assumptions on which the features are based while also discussing some related features that could have been used.

We generally assume that having many players to defend is beneficial for the ball losing team [4]. However, due to the direct and fast nature [10] of counterattacks, it may be difficult for some players that were previously on offense to do so. Thus, we estimate whether a player is able to defend against a possible counterattack by comparing its position with the position of the ball. Therefore, we count all players with less distance to the losing team's goal than the ball and refer to this feature as *losing players behind ball*.

This feature is related to the number of ball winning players behind the ball. These players will be in an ideal position to conduct the counter attack. Furthermore, the ratio or difference between losing and ball winning players behind the ball could also be assessed. Choosing *losing players behind ball* thus emphasises the importance of defending and decelerating counter attacks with numbers over the numerical balance or attacking player number during the counter attack.

Moreover, we presume that a compact defensive shape closes free spaces on the pitch and, thus, limits the possibilities for the attacking team. Vertical compactness is assumed to be especially effective [4]. In its simplest form, a compact shape can be viewed as a dense formation with small distances between players. Thus, we compute the average vertical distance between fieldplayers of the ball losing team and refer to this feature as the *losing vertical compactness*.

The horizontal compactness of the ball losing team is a related feature to its vertical compactness. By focusing on vertical compactness, the assumption is that counter attacks are much less likely to succeed, if the team has moved up the pitch cohesively and has small distances between the players vertically. This usually implies, that the opponent has had to move back as well to defend the cohesive attack.

Finally, we assume that controlling space around the ball is beneficial for defense as it may prevent progress of the ball winning team [4]. We, thus, compute space occupation for every pitch location using the approach from Fernandez et al. [5]. To value spaces around the ball we use an exponentially decreasing function of the ball distance with decay parameter $\alpha = 0.2$. This decay parameter is equivalent to a half-distance of the space control of $3.5m$, which means that a player 3.5 meters away from the ball contributes half as much as a player directly positioned on the ball. We aggregate the space occupation for both teams into a single value by choosing a negative sign for the ball winning team. Thus, we refer to this value as *losing pitch control* since it corresponds with spatial dominance of the losing ball team.

4.2 Influence of Ball Loss Location for Feature Assessment

The ratio of counterattacks (see Figure 1) reveals a strong influence of the ball loss location for counterattack probability. Turnovers in the pitch center have a higher probability to become counterattacks than turnovers closer to the winning team's goal. This is likely caused by the lower distance that needs to be overcome within a given time to count as a counterattack (see Section 3.2).

However, this influence needs to be included in the assessment of features. All three constructed features correlate with ball loss location to some degree. Therefore, they already possess some predictive power that does not stem from their tactical relevance. Thus, we do not evaluate overall counterattack probability but rather the derivation to the average ratio at a location (see Figure 1). Hence, to gain an impression of prediction capability, we employ feature-specific thresholds that were chosen to separate the set of *persistent turnovers* into two comparably large but also distinct subsets. We group all turnovers with feature values below the threshold into one group and all turnovers with values above the threshold into another group. For both subsets, we compute individual counterattack ratios for every pitch location (as presented in Section 3.3). Finally, we respectively subtract the average counterattack ratio to obtain differences.

4.3 Prediction Capability of the Constructed Features

Losing Players Behind Ball We choose a threshold of eight players and yield two comparably sized subsets (see Figure 3). The differences to the average ratios display a largely increased probability for below-threshold turnovers. This effect is especially valid in the pitch center while a low number of turnovers near the corner flags influences the results.

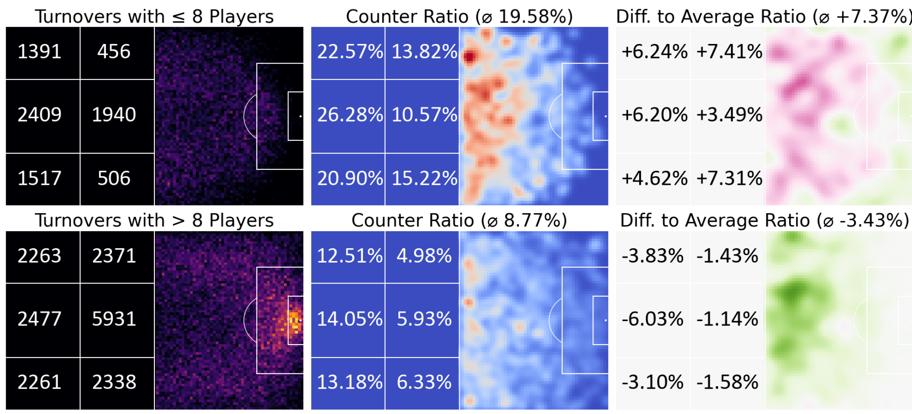

Fig. 3. Influence of *losing players behind ball* for counterattacks. Left: Number of turnovers. Middle: Ratio of counterattacks. Right: Difference to the average counter-attack ratio (Figure 1). Top row: Below-threshold turnovers with eight or less players. Bottom row: Above-threshold turnovers with at least nine players. For the right plot, we choose green for values below and pink for values above location average. The maximum value is ±15%.

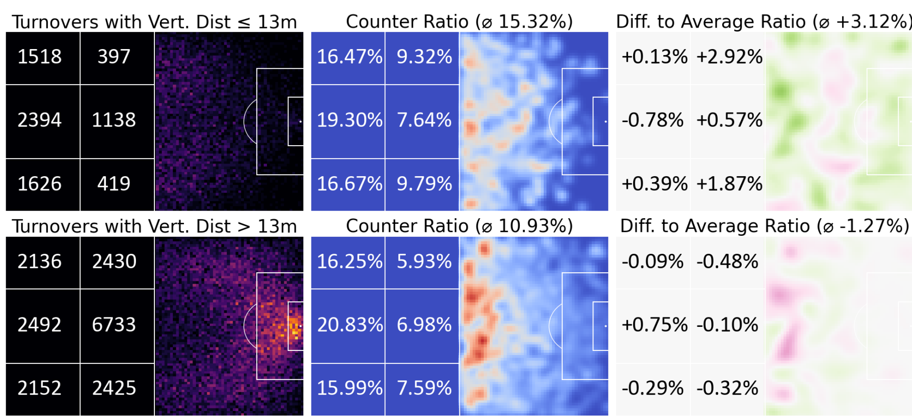

Fig. 4. Influence of *losing compactness* for counterattacks. Top row: Compactness up to 13m. Bottom row: Compactness of more than 13m. For full caption see Figure 3.

Losing Vertical Compactness In this case, we choose a threshold of 13m to create two distinct subsets (see Figure 4). The differences to average ratio show a small zone-dependent effect of *vertical compactness*. The below-threshold turnovers (good *vertical compactness*) have a slightly decreased probability of a counterattack in the pitch center. For zones near the sidelines, however, this effect is not observable. Yet again, the zones near the corner flags show a considerably small number of below-threshold turnovers which might influence the results.

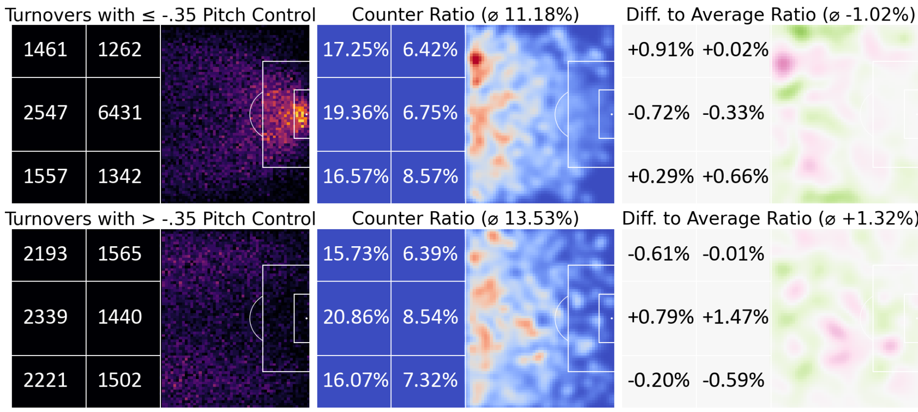

Fig. 5. Influence of *losing pitch control* for counterattacks. Top row: Pitch control of -0.35 or less. Bottom row: Pitch control of more than -.35. For full caption see Figure 3.

Pitch Control around the Ball We create subsets using a feature threshold of -0.35 (see Figure 5). A negative threshold was chosen to obtain comparable sizes as the ball winning team naturally controls more space in the majority of cases. The results indicate ambiguous influence of pitch control on counterattack probability. Above threshold turnovers (good *losing pitch control*) show a slightly decreased probability of a counterattack for zones near the sidelines. However, they also show a slightly increased probability of a counterattack at the pitch center. This is likely because good occupation of spaces around the ball correlates with the poor occupation of spaces away from the ball. Thus, this effect might describe failed pressing attempts of the ball losing team to some degree.

5 Model-based test of features

We test our set of comprehensible features in a predictive logistic regression model with balanced class weights. To incorporate the influence of location we use the direct distance to the losing team's goal additionally to the three constructed features. A binary Logistic regression with 'l2' regularization was performed using the handcrafted features (see Table 1) as independent variables. The model was further evaluated using five-fold cross-validation which was performed in order to check the out-of-sample validity. Furthermore, feature coefficients and their respective odds ratios were computed to assess the influence of the features. The results show that losing players behind ball had the highest (negative) influence on counterattacks. Pitch control shows the second-highest influence in predicting a successful counterattack. The goal distance and losing vertical compactness show negligible effects. The results of the 5-fold cross validation display a general capability of the model to predict counterattacks. Small deviance between individual splits indicates the stability of the model. However, the absolute values still show room for improvement.

Table 1. Feature influences and results of the logistic regression model. Left: Feature coefficients and odds ratio (OR). Right: Cross-validation results for different metrics.

Feature	Coeff.	OR
Intercept	4.285	72.591
Los. Players Behind Ball	-0.218	0.804
Losing Pitch Control	-0.150	0.860
Los. Team Goal Distance	-0.027	0.973
Losing Vert. Compactness	-0.008	0.993

Metric	Cross-val results
F1-Weighted	0.6417 ± 0.0029
AUC	0.6916 ± 0.0037
Accuracy	0.6418 ± 0.0029
Recall	0.6532 ± 0.0077

6 Conclusion

We have made a first approach towards an exhaustive expected counter metric. Using our proposed framework for understanding complex sequences we have created a dataset of *sequences of interest* with counterattacks labels based on intuitive *success criteria*. We have constructed three comprehensible features for counterattack success and assessed the prediction capability of the features using manually defined threshold values. Based on the features we created a simple predictive logistic regression model. The framework has revealed various comprehensible insights in how to prevent counterattacks. Most prominently, players behind the ball are highly important for preventing counterattacks. Thus, players in the back need to avoid turnovers (e.g. during build-up play) as much as possible. More space control is also beneficial to prevent counterattacks, especially at the sidelines. Less significant, a vertical compact shape is beneficial in the center. Our model predicts the risk of a turnover with substantial performance. Thus, it can be employed for various sport-specific applications (e.g. evaluate tactical fouls or create risk-reward profiles). Admittedly, the assessment with hand-defined thresholds does not capture fine-grained statistical anomalies in feature distributions. Moreover, our model depends on choices influencing the dataset (*sequences of interest*) and features. Thus, in future we plan to carefully evaluate these choices and to incorporate more sophisticated continous *success criteria* [6]. Finally, our approach may benefit from a larger dataset. However, due to the detailed insights gained on the matter of counterattacks we promote our framework for analyzing all types of complex sequences in soccer (e.g. build-up play or different set pieces).

Acknowledgements This research was supported by a grant from the German Research Council (DFG) to DM (grant ME 2678/30.1).

References

1. Anzer, G., Bauer, P.: Expected passes: Determining the difficulty of a pass in football (soccer) using spatio-temporal data. Data Mining and Knowledge Discovery **36**(1), 295–317 (Jan 2022). https://doi.org/10.1007/s10618-021-00810-3, https://link.springer.com/10.1007/s10618-021-00810-3

2. Bauer, P., Anzer, G.: Data-driven detection of counterpressing in professional football: A supervised machine learning task based on synchronized positional and event data with expert-based feature extraction. Data Mining and Knowledge Discovery **35**(5), 2009–2049 (Sep 2021). https://doi.org/10.1007/s10618-021-00763-7, https://link.springer.com/10.1007/s10618-021-00763-7
3. Bauer, P., Anzer, G.: A goal scoring probability model for shots based on synchronized positional and event data in football (soccer). Frontiers in Sports and Active Living **3**, 53 (2021). https://doi.org/10.3389/fspor.2021.624475, publisher: Frontiers
4. Fernandes, T., Camerino, O., Garganta, J., Pereira, R., Barreira, D.: Design and validation of an observational instrument for defence in soccer based on the Dynamical Systems Theory. International Journal of Sports Science & Coaching **14**(2), 138–152 (Apr 2019). https://doi.org/10.1177/1747954119827283, http://journals.sagepub.com/doi/10.1177/1747954119827283
5. Fernandez, J., Bornn, L.: Wide Open Spaces: A statistical technique for measuring space creation in professional soccer. In: Sloan sports analytics conference. vol. 2018 (2018)
6. Fernández, J., Bornn, L., Cervone, D.: A framework for the fine-grained evaluation of the instantaneous expected value of soccer possessions. Machine Learning **110**(6), 1389–1427 (Jun 2021). https://doi.org/10.1007/s10994-021-05989-6, https://link.springer.com/10.1007/s10994-021-05989-6
7. Groll, A., Schauberger, G., Tutz, G.: Prediction of major international soccer tournaments based on team-specific regularized Poisson regression: An application to the FIFA World Cup 2014. Journal of Quantitative Analysis in Sports **11**(2) (Jan 2015). https://doi.org/10.1515/jqas-2014-0051, https://www.degruyter.com/document/doi/10.1515/jqas-2014-0051/html
8. Hewitt, A., Greenham, G., Norton, K.: Game style in soccer: what is it and can we quantify it? International Journal of Performance Analysis in Sport **16**(1), 355–372 (2016)
9. Hockeyviz: Smoothing: How to (2022), https://hockeyviz.com/howto/smoothing
10. Lago-Ballesteros, J., Lago-Peñas, C., Rey, E.: The effect of playing tactics and situational variables on achieving score-box possessions in a professional soccer team. Journal of Sports Sciences **30**(14), 1455–1461 (2012), publisher: Taylor & Francis
11. Liu, G., Luo, Y., Schulte, O., Kharrat, T.: Deep soccer analytics: learning an action-value function for evaluating soccer players. Data mining and knowledge discovery **34**(5), 1531–1559 (2020)
12. LLC, S.: Playing Styles Definition by StatsPerform (2022), https://www.statsperform.com/resource/stats-playing-styles-introduction/
13. Memmert, D., Raabe, D.: Data Analytics in Football: Positional Data Collection, Modelling and Analysis. Routledge, Abingdon, Oxon ; New York, NY : Routledge, 2018., 1 edn. (May 2018). https://doi.org/10.4324/9781351210164, https://www.taylorfrancis.com/books/9781351210157
14. Raudonius, L., Allmendinger, R.: Evaluating Football Player Actions During Counterattacks. In: International Conference on Intelligent Data Engineering and Automated Learning. pp. 367–377. Springer (2021)
15. Robberechts, P., Van Haaren, J., Davis, J.: A Bayesian Approach to In-Game Win Probability in Soccer. Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining pp. 3512–3521 (Aug 2021). https://doi.org/10.1145/3447548.3467194, http://arxiv.org/abs/1906.05029, arXiv: 1906.05029

16. Spearman, W.R., Basye, A.T., Dick, G.J., Hotovy, R., Hudl, P.P.: Physics-Based Modeling of Pass Probabilities in Soccer (2017)