Beyond Expected Goals

William Spearman
Liverpool Football Club

MIT SLOAN
SPORTS ANALYTICS CONFERENCE
presented by ESPN

Beyond Expected Goals

William Spearman
Hudl
william.spearman@hudl.com

Abstract

Many models have been constructed to quantify the quality of shots in soccer. In this paper, we evaluate the quality of off-ball positioning, preceding shots, that could lead to goals. For example, consider a tall unmarked center forward positioned at the far post during a corner kick. Sometimes the cross comes in and the center forward heads it in effortlessly, other times the cross flies over his head. Another example is of a winger, played onside, while making a run in past the defensive line. Sometimes the through-ball arrives; other times the winger must break off their run because a teammate has failed to deliver a timely pass. In both circumstances, the attacking player has created an opportunity even if they never received the ball. In this paper, we construct a probabilistic physics-based model that uses spatiotemporal player tracking data to quantify such *off-ball scoring opportunities* (OBSO). This model can be used to highlight which, if any, players are likely to score at any point during the match and where on the pitch their scoring is likely to come from. We show how this model can be used in three key ways: 1) to identify and analyze important opportunities during a match 2) to assist opposition analysis by highlighting the regions of the pitch where specific players or teams are more likely to create off-ball scoring opportunities 3) to automate talent identification by finding the players across an entire league that are most proficient at creating off-ball scoring opportunities.

1. Introduction

How can we quantify the value of a player standing, unmarked, at the far post waiting for a cross to come in that never arrives? Alternatively, how can we quantify the value of a striker who has positioned himself in space near the penalty spot but gets nothing out of it because the winger fails to deliver the square pass? The work we present in this paper endeavors to answer these and other questions related to the effect of off-ball positioning on the probability of scoring.

Compared to many other sports, scoring in soccer is a rare phenomenon. Because of this, various statistics are used as a proxy for the performance of a team throughout the course of match. Simple counting statistics such as the number of shots taken, the number of crosses, or the number of passes into the box are used to judge team performance. In 2012 a new metric, expected goals [1], was developed by Sam Green to quantify the probability of a shot resulting in a score. Various non-shot expected goal extensions [2] [3] [4], have been proposed that use non-shot events to quantify the likelihood that a given sequence of play will result in a score.

With the proliferation of spatiotemporal tracking data, exciting new ways of measuring the probability of scoring have been developed by Patrick Lucey et. al. [5], Daniel Link et. al. [6], and others. The research in [5] quantifies the probability that a shot will result in a score using strategic features from the 10-seconds of play before the shot; the *dangerosity* metric introduced in [6] represents an innovative heuristic for quantifying the scoring danger posed by a scenario with explicit modeling of important factors such defensive pressure and the number of attackers and defenders in key regions in front of the goal.

In this paper, we attempt to build a model that represents the *probability* that a player not currently in possession of the ball will score based only on the *instantaneous* game state. We term this the *off-ball scoring opportunity* (*OBSO*).[1] There are three aspects about our approach that we would like to highlight:

- **Off-Ball Chance Creation:** Players can be given credit for creating space in a scoring location even if the ball is never delivered to them.
- **Understandable Modeling:** Each component of the model has real soccer meaning and answers a specific soccer question such as: "where will the next pass go?" or "which regions of the pitch are controlled by the attacking team?". Many of these model components could form the basis for additional analysis with tracking data to answer questions about soccer phenomena apart from scoring.
- **Scoring Predictiveness:** As with a traditional expected goals model, the *OBSO* can be integrated to yield the total expect score. This allows us to compute the expected scoring production for a player or team or during the course of a match. In addition, due to the spatial aspect of the model, we can also predict *where* on the pitch and *when* during the match goals are most likely to have occurred.

At the core of our model is an extension to the *pitch control field* first developed in [7] that quantifies the *potential* regions of control at some time in the near future. The *potential pitch control field* and derived models provide a new framework for interrogating the tracking data to answer questions about the short-term evolution of the game-state.

2. Data

For this analysis we use soccer match data produced by Hudl. The data spans 58 matches played between teams from a 14-team professional soccer league during the 2017-2018 season. Our dataset includes both event data and spatiotemporal tracking data.

The event data comprises the on-ball actions that were performed during the match. For each event, the following information is known: 1) the match time at which the event occurred 2) the on-ball player 3) the type of event (e.g. pass, shot, goal, etc.). The tracking data represents the location of every player on the pitch with a temporal frequency of 25 Hz and the corresponding match time for each tracking frame is specified.

2.1. Data Preprocessing

Our goal is to produce an analysis that is relatively insensitive to data source, data quality, and requires minimal data processing and cleaning. Accordingly, we perform only two steps when processing the data: 1) tracking data smoothing 2) naïve data synchronization.

[1] For this paper, we will often refer to *off-ball scoring opportunity* as *opportunity*. If the word *opportunity* is used in a more general context, it will not be italicized.

To smooth the tracking data, we perform a least-squares smoothing [8] that reduces noise and computes the instantaneous values of the higher order derivatives: velocity and acceleration. As a convention, we rotate the data by 180° when needed to ensure that the home team is always attacking from left to right. When data from both home and away teams is displayed, the home team is shown in red while the away team is shown in blue.

The moment data is synchronized with the tracking data by matching the event timestamp with the match time of the corresponding tracking frame. This allows us to take snapshots of the match for each event that let us know who is interacting with the ball, how they are interacting with the ball, and the locations of all the other players on the pitch at that point during the match. The event data timestamps have some level of stochastic noise which leads to a temporal uncertainty on the order of three seconds. This can lead to moderate errors in the positions of the players. To minimize the effect of these errors, we choose a Bayesian approach to fitting our model (see Section 3.5 for a full description of our parameter estimation procedure) and select strong priors based on independently measurable/known information.

2.2. Data Anonymization

The data used for this analysis has been anonymized and the teams are referred to by a letter from A-N. Although only 58 games are used for the analysis, mid-season league standing of the teams are used to describe the team using a single letter. Thus, the team with the best mid-season record is Team A while the team with the worst mid-season record is Team N. In addition, player jersey numbers visible in certain visualizations have also been randomized and do not correspond to the true jersey number of the player.

2.3. Reproducibility

In an effort to maximize analysis reproducibility, modeling decisions are made to minimize the impact of data inaccuracies that may vary between providers of spatiotemporal tracking and event data. Ball data is not needed as on ball-events can be used to identify which player is currently touching the ball. The minimum event data required is: 1) when the event occurred 2) who touched the ball during the event and 3) whether the event was a goal, a shot, or another on-ball event. Although the tracking data used for this analysis was collected at 25 Hz, it is only necessary to have a snapshot of the positions and velocities of the players for each on-ball event. This reduces the required frames from ~8,000,000 (25 Hz for 90 minutes) to a more manageable ~1000 (one per event).

3. Methods

To compute the probability that an attacking team will score with the next on-ball event, we must compute the probability of a pass to each of the other 10 players and the probability that the receiving player will shoot and score. Of course, some of these probabilities are bound to be very small, e.g. the probability of the keeper scoring is almost always zero even though the back pass to the keeper is likely. Although conceptually simple, it is difficult to compute the above transition probabilities without spatial information. To simplify, we rephrase the problem as follows: what is the probability that the attacking team successfully passes to each point on the pitch and scores from that location. This can be conceptualized as a three step process.

- **Transition:** Probability that the next on-ball event occurs at an arbitrary point, *r* on the pitch. Denoted by: T_r.
- **Control:** Probability that a ball at point *r* will be controlled by the passing team. Denoted by C_r.
- **Score:** Probability of scoring from point *r* on the pitch assuming the next on-ball event occurs there.[2] Denoted by G_r.

The total probability of scoring with the next on-ball event for the attacking team can therefore be described by the summation in Equation 1:

$$P(G|D) = \sum_{r \in \mathbb{R} \times \mathbb{R}} P(G_r \cap C_r \cap T_r|D) \tag{1}$$

In this equation, *D* represents the instantaneous state of the game (player positions and velocities). The conjoined probability can be decomposed into a series of conditional probabilities as follows:

$$P(G|D) = \sum_{r \in \mathbb{R} \times \mathbb{R}} P(S_r|C_r, T_r, D)P(C_r|T_r, D)P(T_r|D) \tag{2}$$

In the next sections, we will address each term and the simplifications we use to compute them.

3.1. Control Model: The Potential Pitch Control Field

The most difficult expression in the decomposed conditional probabilities presented in Equation 2 is the second term: $P(C_r|T_r, D)$, the probability that the attacking team will control the ball at point *r* assuming the next on-ball event occurs there. We introduce a new model, the *potential pitch control field (PPCF)*. This concept borrows heavily from the *pitch control field* developed in [7]. Both models assume that while in proximity to the ball, a player's ability to make a controlled touch on the ball can be treated as a Poisson point process. The longer a player is near the ball without another player interfering, the more likely it becomes that they are able to make a controlled touch on the ball. As in [7], we seek to quantify the probability of control for each player at each location on the pitch. Instead of assuming the ball were to arrive at the destination instantaneously, we compute the time it would take the ball to arrive using aerodynamic drag. This means that more distant passes will give defenders and attackers more time to converge on the specified position. Furthermore, we introduce a decision making aspect to the model whereby the pass is assumed to travel at the most advantageous physically realistic time for the attackers. Before considering the full details of this model, an illustrative example of the *PPCF* is shown in Figure 1 below.

[2] Note: this does not assume that a shot is taken, so this term encompasses the probability that the on-ball event is a shot and the probability of that shot being successful.