Sport Analytics: Modeling the Significance of Soccer Passes

By Michael Liang and Liam Sharp

1. Introduction

A critical element of soccer is the action of passing the ball, yet traditional statistics poorly quantify them despite the frequency of their occurrence. Though conventional metrics can depict an adequate portrait of a team's passing throughout a match and/or season, measuring the true influence of a team's ball movement is practically impossible without advanced data collection and analytics. For instance, inferring that Manchester City is the best passing team in the Premier League because they lead the field in average passes per game (688.39) could be a misleading conclusion. [1] The statement may very well be true, but relying on traditional statistics, which by nature omit many crucial aspects of the sport, is a risky proposition when it comes to evaluating performance and developing strategies. Consequently, soccer clubs that are not invested in data analytics run the risk of being left behind by their competitors.

This paper aims to develop a method of evaluating soccer passes throughout a match. We will review a rudimentary "impact factor" method and propose our own system for evaluating passes. In the conclusion, the applications and benefits of our method, as well as drawbacks and how it could be improved will be discussed in further detail.

2. Literature Review

2.1 The "Impact Factor"

A proposed method of pass evaluation in the description of the project is by assigning a score to every pass based on the number of "outplayed" opposing players in the x-direction of the soccer pitch. For example, a pass that outplays three opponents will be awarded a score of 3, while a pass that outplays no opposing players will score a 0. In this model, a high number suggests a pass is impactful, and a low number signifies a pass with negligible influence.

The impact factor generated from this method can be useful in evaluating some aspects of a team's passing game. Tracking how many opponents are outplayed per pass can be a useful indicator of a team's willingness to take risks and move the ball quickly across the pitch. In addition, the impact factor can reveal the passing efficiency of a team and/or player. Suppose a team has many passes with a score of 0-1. In that case, a reasonable conclusion could be that the team in question is playing too conservatively at the expense of generating real scoring opportunities. If a team has a large quantity of passes with an impact factor ranging from 3-5, perhaps we could infer that they are more willing to sacrifice possession of the ball for a potential chance at the goal.

Despite its uses, the impact factor model does have its flaws, particularly in its bias towards offense. Maintaining possession of the ball by passing backward is optimal in many situations. However, without proper context evaluation, crucial but conservative passes are not properly represented. The model also weights every pitch area equally, when in actuality, passes completed near either extremity of the pitch are significantly more meaningful than the passes that occur near the center of the pitch. In general, the impact factor can deduce some information but does not accurately encompass all areas of the passing game.

2.2 Expected Passes by Anzer and Bauer

Anzer and Bauer use spatio-temporal tracking data to measure the difficulty of passes in soccer. The machine learning model improves the conventional pass performance model, which employs a binary pass completion metric by adding context to individual passes through an extremely detailed data set. The resulting probability of any given pass being completed is a better indicator of a team's passing quality than the non-contextual completion metric. [2]

In our proposed model in the following section, we attempt to contextualize every pass using the data at our disposal. We quantify the difficulty of passes in the form of generated pressure from surrounding opponents to a pass (openness of both the passer and recipient). We also partition the soccer pitch into zones of equal size and allocate differing weights to each area based on the average scoring probability of that area. Though Anzer and Bauer are primarily concerned with pass completion probabilities, there is a clear overlap with our method.

3. Data

The data in this paper was taken from Metrica Sports open sample data set. The data is anonymized, meaning there are no references to the players' names, club names, and their competitors. Tracking and event data are synchronized, and data was taken from the sample match 1 and 2 database.

4. Methodology

Our impact factor considers four main characteristics to minimize bias and maximize statistical accuracy. The pass characteristics we examined are openness, the type of pass, possession (based on the outcome of the pass), and expected threat based on the origin and destination zone of the pass. More reasoning behind our methodology will be provided in section 5.

Fundamentally, passing has two primary functions in soccer.

a) Maintaining possession of the ball.
b) Generating scoring opportunities.

The four factors we have chosen to emphasize for calculating our impact factors ensure we evaluate both aspects of the passing game. If, for instance, all of our factors were biased towards exclusively (a) or exclusively (b), the resulting impact factor would not be an excellent indicator of the overall passing of a player and/or team. Our method evaluates two factors that favor objective (a) (Possession and Openness) and two factors that favor objective (b) (Type of pass and Generated Expected Threat). This improves upon the initial impact factor method presented in class by considering more data in the provided set, such that it encompasses all elements in the overall passing game. The resulting impact factor of any given pass will be the sum of its possession value and attacking value.

4.1 Openness

In order to evaluate the openness of passes, we look at the pressure generated from the opposing team to the player who is passing the ball, as well as the recipient of the pass. Pressure is computed using two measurements.

a) The closest opponent to the passer and recipient is considered in meters. Passing to a teammate who is facing minimal pressure will increase the possession value of a pass. The value can also increase if the player who completes the pass can do so while facing heavy defensive pressure.

b) The total number of opposing players around both the passer and receiver within a radius of 10 meters. The possession value score increases logarithmically as the number of opponents surrounding the pass increases.

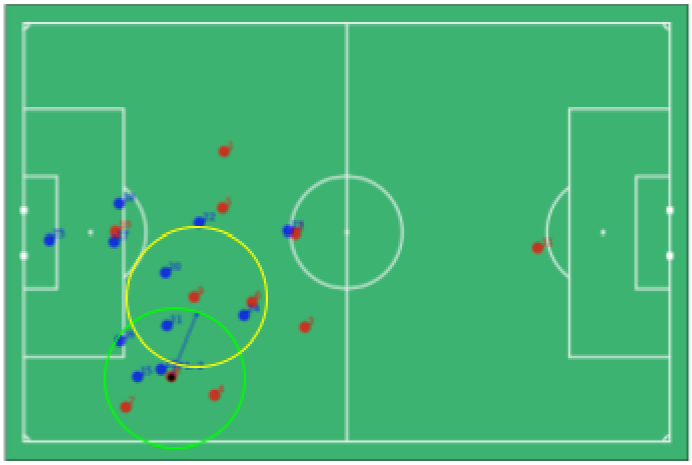

Fig.1. *A visual representation of the 10 meter radius between the passer (green) and receiver (yellow). In this example, there are four defenders surrounding the passer and three defenders around the recipient.*

4.2 Type of pass

In order to classify passes, we consider only the passes which occur in the attacking 60% of the pitch. A passer can be outside of this area of consideration provided the destination of the pass is in the attacking zone. We chose to filter passes in the defensive zone (40%) because they are often possession passes with little pressure. In the analysis section, we will discuss some of the drawbacks of this filtering process in more detail.

Based on the aforementioned definition, we separated the eligible passes into three categories.

a) Progressive: Passes in the attacking 60% of the pitch that travel 25% closer to the goal from the starting pass point. Attributed weight in the cumulative impact factor is **0.6.**
b) Neutral: Passes in the attacking 60% of the pitch that travel between 0% to 25% closer to the goal from the starting pass point. Attributed weight in the cumulative impact factor is **0.5.**
c) Defensive: Passes that are neither progressive nor neutral. Attributed weight in the cumulative impact factor is **0.4.**

4.3 Possession

We measured possession by looking at the resulting outcome of every series of passes. We define a series as consecutive uninterrupted passes. A new series of passes is created whenever there is a challenge on the ball or a change of possession. The sequence of passes is given a score based on the result of the play.

- If a series leads to a goal/shot, then its corresponding score will be **1.**
- If a series leads to a corner, free kick, or throw-in (the team with the ball must maintain possession), then its corresponding score will be **0.5.**
- If a series leads to a lost possession, then its corresponding score will be **0.1.**

4.4 Generated Expected Threat

We reference the model presented in Introducing Expected Threat (xT). The pitch is divided into 96 equally-sized zones, each with their respective weight threat based on their expected goals (xG) scores. Expected goals are a performance metric that allocates a probability that a team generates a scoring opportunity to all areas of a soccer pitch. Expected Threat builds upon xG but also values locations on the pitch based on its potential to induce danger later in the possession. [3]

We can evaluate how effective a pass is at generating scoring opportunities by looking at its generated expected threat, which is quantifiable by computing the difference in xT scores between the destination and origin of the pass.

$$xT_{pass} = xT_{end} - xT_{start}$$

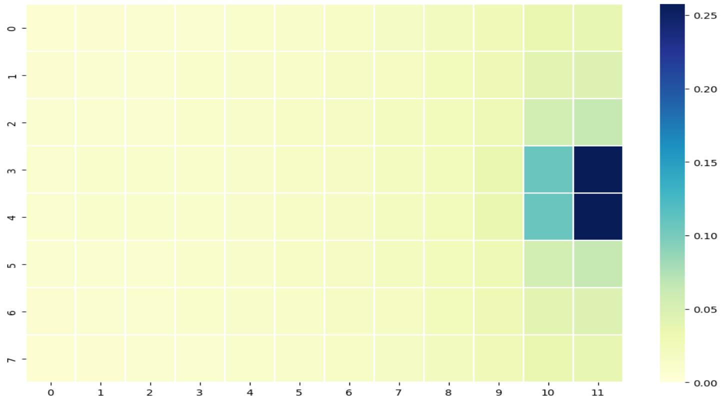

Fig. 2. *Soccer pitch divided into xThreat zones. Most areas are considered low-threat apart from areas closer to the opponent goal. Attacking direction is towards the right side of pitch.*

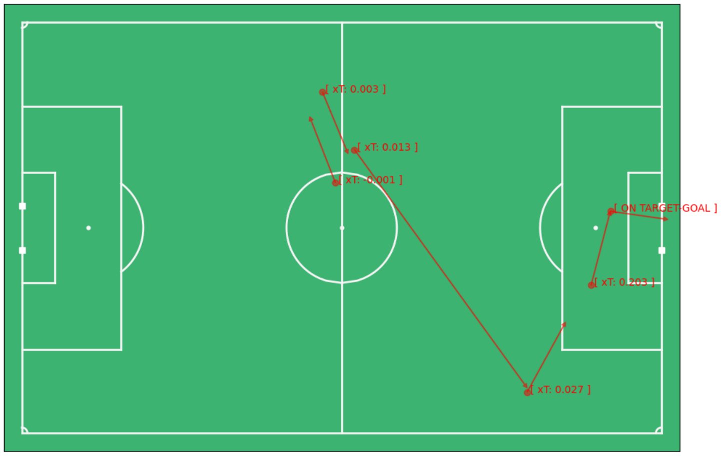

Fig.3. *The Expected Threat generated from passes. Passes that move toward the opponent's goal (right) will take on positive values.*

5. Analysis and Applications

In order to deduce a cumulative impact factor for the passes in the data set, we propose two methods. The limitations of each method will be presented in section 6.

a) Computing the weighted average of the individual factors, in which each carries an equal weight of 25 per cent.

b) Computing a revised weighted sum of the factors. Every pass is assigned a value for its possession and attacking influence based on the four individual factors. Values are weighed by the type of the pass, and the summation is a pass' cumulative impact factor.

5.1 Weighted Average

The cumulative impact factor for each individual pass in the data set is calculated by taking the weighted average of a pass' four factors, each carrying an equal weight of 25 per cent. The decision to weigh every factor evenly ensures we minimize bias in the analysis. Evaluating four different factors ensures the accuracy of the resulting impact factor by adding context that will minimize the chance of fluke errors in the results.

5.2 Revised Weighted Sum

The factors we computed in our methodology can be connected to create a general cumulative impact factor for each pass using the following formula:

$$Revised\ Weighted\ Sum = (1 - w_i)[1 - \lambda\, exp(\frac{-(x_p + x_r)}{2})] + w_i\left[1 - 0.8\, exp(-2\mu)\right]$$

- ➢ $w_i = weight\ associated\ with\ pass\ type.\ \ for\ progressive\ (0.6),\ neutral\ (0.5),\ and\ defensive\ (0.4).$
- ➢ $x_p + x_r = The\ sum\ of\ all\ nearby\ opponents\ near\ the\ passer\ (x_p)\ and\ recipient\ (x_r).$
- ➢ $\lambda = 1/(1 + y),\ where\ y = possession\ outcome\ score.\ Refer\ to\ section\ 4.3$
- ➢ $\mu = xT_{pass}(speed_{pass}/AvgSpeed_i),\ where\ AvgSpeed_i = Average\ speed\ of\ the\ pass'\ type$

5.2 (continue) The first component of the sum describes the possessive value of a pass, and is measured using the possession outcome factor and the sum of all opponents within a 10 meter radius of the passer and receiver. The second component of the sum describes the attacking value of the pass, and uses the xThreat generated by the pass and the pass' speed relative to its type's average. Weighing each component by its pass type ensures that defensive/conservative passes are not being incorrectly judged. The nature of the formula allows for progressive passes to have more weight associated with its attacking value, while defensive passes have more weight with the pass' progressive value. A couple of technical caveats to the formula ensure that its results are as intended.

a) The coefficient 0.8 in the second component shifts the function upwards so that if μ is negative, the attacking value can still take on a positive value (mainly applies to defensive passes).
b) Dividing the expression $-(x_p + x_r)$ in the possessive value by 2 ensures that the sum of surrounding players doesn't exceed a threshold that makes the exponential function insignificant.
c) Each component is of the form $1 - exp(-[function])$ in order to maintain an output that is between 0 and 1. This form also supports the notion that defensive pressure does not linearly increase as the number of defenders increases. In reality, the pressure magnitude grows logarithmically as more and more defenders are added to the mix.

The cumulative impact factor we obtain using this formula will compute a result ranging from 0 to 1. A score of 1 would signify that a given pass was as effective as can be. We attempt to generate an impact factor that can discern that neutral and defensive plays are often critical to a soccer match's progression.

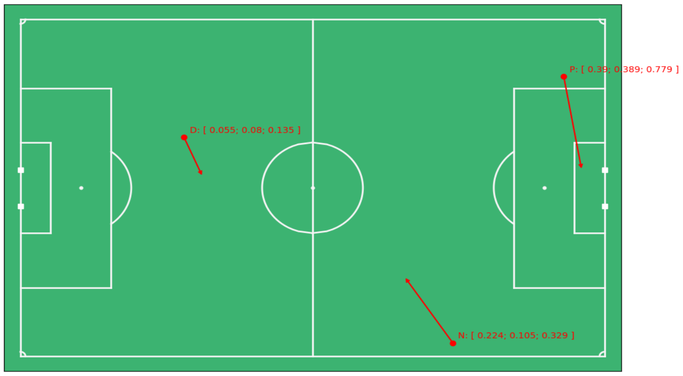
P: [0.39; 0.389; 0.779]
D: [0.055; 0.08; 0.135]
N: [0.224; 0.105; 0.329]

Fig. 4. *The cumulative impact factor for three different types of passes where the attacking direction is towards the right side of the pitch. Each pass has a corresponding possessive value (first component of vector) and attacking value (second component) that make up its cumulative impact factor (third component).*

Both methods have many applications at the player and team levels. Individual players can refer to each factor that makes up their cumulative impact score to see where they can improve. While the overall score may not be useful for a player on the surface, they can still focus on specific aspects in great detail due to our impact score being composed of four primary elements. Teams can use computed analytics to indicate better the efficiency and methods in which their club handles the ball. Strategically, a club can evaluate the success rate of a particular set play, or uncover what types of plays they are susceptible to when they are on defense.

6. Conclusion

The objective of our model is to evaluate the value of passes in a soccer match. Using tracking and event data, we computed a metric that considers the openness of the pass, its type, the resulting possession, and its generated expected threat. Factors can be evaluated individually to evaluate specific areas of a team/player's passing game, or more generally by referencing the cumulative impact factor of the passes. We provide two possibilities for calculating a single score for all passes, either by taking the weighted average of our four computed factors, or by computing the weighted sum using the formula presented in section 5.2.

Our impact factor improves on the proposed model in the outline of the project by adding further context using more data in the provided set. By increasing the number of factors that make up our cumulative score, we have minimized the bias in the analysis towards offense, in order to provide a more accurate depiction of the sample team's overall passing influence.

6.1 Limitation and Looking forward:

Some notable limitations to our pass evaluation model lies in the data collection itself. Namely, incomplete passes are omitted. Even though a negative outcome transpires (in this case, a turnover), data analytics suggests that judging passes solely on whether they are completed or not is an archaic way of evaluating a team's passing offense. There are instances when theoretically correct decisions (probability-wise) do not translate to positive outcomes, which doesn't mean the decision was incorrect. Ideally, teams should evaluate all passes in a game, regardless of their outcome, and not solely those that complete.

For the weighted average model, four factors comprise each pass to compute a cumulative impact score. A better (albeit less succinct) way to evaluate the passes would be to keep the factors separate and simply apply the individual scores to relevant contexts. The cumulative score is an attempt to summarize a pass as generally as possible given our methodology, but analyzing the specific factors of the passing game individually would likely be a more accurate method of studying passes in the long run.

In the weighted sum method, we assign specific weights to certain factors based on the data-provided context of the pass. In the possessive component of this model, we do not consider the computed values from section 4.1(a) in our cumulative impact factor analysis. We only included the total number of players within the proximity of the pass in the calculation.

Another limitation in the weighted sum assessment is how pass types are weighed in our calculation. The decision to attribute the weights we selected was for the sake of simplicity. Ideally, weights should be derived using data-driven analytics, not for the sake of convenience.

By categorizing types of passes in our methodology, we filter the data such that the passes who's destinations are in the attacking 60 per cent of the pitch. This inherently has a slight bias towards offense, and doesn't track the passes that occur in the defensive zone that are crucial within the context of the game. For example, a pass that dissuades a dangerous scoring opportunity for the opposition is undoubtedly highly-influential, but our current model lacks the ability to discern a meaningful defensive pass from a negligible one.

References

[1] Premier League average passes per game overview for teams. FootballCritic. (n.d.). Retrieved October 5, 2022, from https://www.footballcritic.com/premier-league/season-2019-2020/passes-per-game/2/21558

[2] Anzer, G., &amp; Bauer, P. (n.d.). Expected Passes. Springer. Retrieved October 4, 2022, from https://link.springer.com/content/pdf/10.1007/s10618-021-00810-3.pdf

[3] Singh, K. (n.d.). Introducing expected threat (XT). Karun Singh. Retrieved October 4, 2022, from https://karun.in/blog/expected-threat.html#visualizing-xt