Analyzing Soccer Using Mathematical Models
EE 546 - Final Project

Philipp Schauer

1 Introduction

Almost 20 years ago, the Oakland Athletics - a Baseball team from Oakland - invented a system that was called "Moneyball". The team struggled constantly with a low budget and started to use mathematical models to sign players who would greatly outperform their salary. These models used simple statistics, most of which could be solved by High School students. However, most of the other teams used the so-called "eye test" in determining which players would be successful. So moving away from subjective evaluations to objective models can give a tremendous advantage. Scouts with antiquated concepts were replaced by college graduates in mathematics and economics who have never played the sport but rather have an understanding of statistics. [Lew04]

What was seen skeptical by other teams and even laughed at has since transformed the way sports executives make decisions in different sports. Almost every team in the four major US sports employ data scientists or mathematicians. Nowadays, much more advanced models including Machine and Deep Learning algorithms are involved in order to make more informed decisions and have a competitive advantage over the other teams. It is self-evident that wherever a lot of data is involved, you can use this data to gain an advantage. This can include ticket pricing, tactics for games and signing players.

2 Motivation

While other fields allow for exact models, using data science causes problems in sports. You can fit an image processing algorithm with as many images as you like but in sports you can only use the games that have actually been played, so the sample size is often quite small. Also, human influences sometimes can not be evaluated in numbers. Maybe a player can play well with one team but not in another one because he does not get along with his teammates or simply does not feel comfortable in the new city.

Thus, data scientists need to be very creative with their models so that they can actually help their teams win games. While sports executives in Baseball or American Football have been successful with this, it is much more complicated in Soccer.

This literature review should evaluate a number of research papers about potential ways to use mathematical models in soccer and conclude whether statistical analysis can be used to make the 'Immeasurable Sport' [FBC19] measurable.

3 Expected Possession Value

Soccer is the sport where one score can change the outcome of a match the most. In Basketball, for example, often over 100 scores are made per game, while in soccer sometimes one or two goals are scored in the entire game. This highlights the importance of possessions that end up in a score. Thus, a model that can determine the quality of a possession is a good start to analyze this sport.

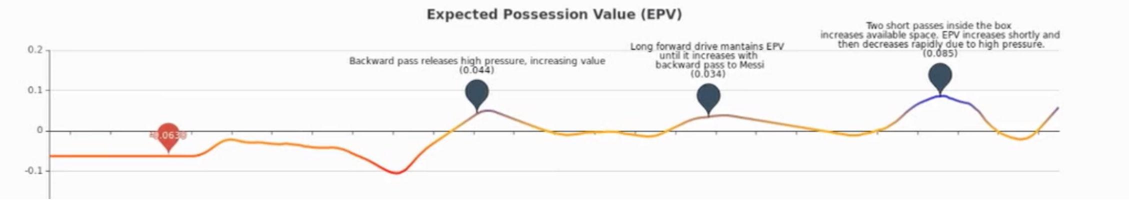


Figure 1: Development of EPV throughout one possession

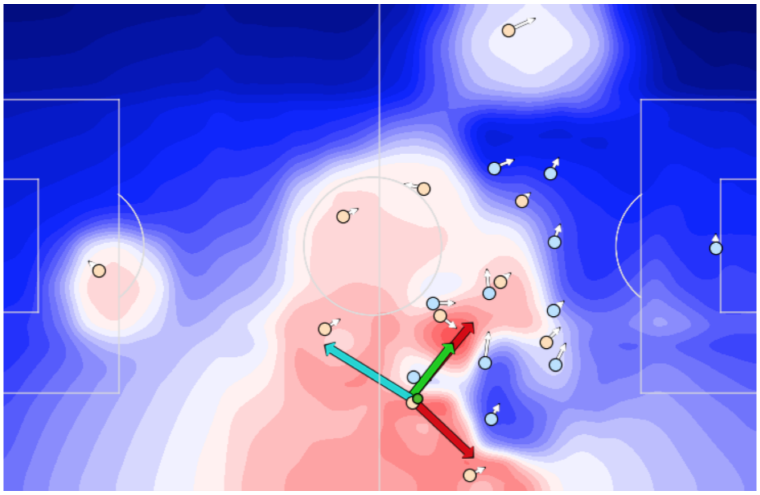


Figure 2: Contour plot of development of EPV given a certain pass direction

3.1 Introducing EPV

The model developed by Fernández et al. is the so called Expected Possession Value. [FBC19]

$$EPV(t) = E[X|T_t] \tag{1}$$

This expression is the expectation of X at a certain time t, where scoring a goal results in $X = 1$, conceding a goal results in $X = -1$, and no goal results in $X = 0$. It is computed as the conditional expectation given a certain state T_t at time t that is modelled as a Markov Decision Process. One state is simply the positioning and trajectory of the players and the ball in x-, y-coordinates at time t. Naturally, a team wants to maximize EPV when it has possession.

The model uses three different actions A that a player can take as the next step:

- **Pass**: $A = \rho$. A pass is when a player tries to pass the ball to one of his teammates.
- **Shot**: $A = \xi$. This action means that the player attempts to score a goal.
- **Ball-Drive**: $A = \delta$. When a player keeps the ball, this is recorded as a ball-drive.

Using the law of total expectations, EPV can now be written as:

$$EPV(t) = E[X|A = \rho] * P(A = \rho) + E[X|A = \xi] * P(A = \xi) + E[X|A = \delta] * P(A = \delta) \tag{2}$$

An example of EPV over the course of one possession can be seen in Figure 1. A good pass leads to an increase in EPV, while a bad pass risks a score by the opponent. This opens a wide range of possibilities. From identifying which players often make the right decisions to finding out which passes create a higher chance of scoring, it can be a big advantage to a team. Figure 2 shows how the value for EPV can change if the player (in the bottom-middle of the field) passes into a certain area.

3.2 Computing EPV

It is difficult to find information on how exactly the different components of the model are computed since the authors of the paper work for sports teams themselves, thus they do not want to lose their competitive advantage by publicly explaining their methods. However, they gave a few hints on how they did that.

As a main tool, the model uses image processing, where passes were analyzed and given a label (*success* or *no success*). The following tools were also used:

- **Machine Learning** is used in computing the expectation of X given that the next action is a pass. Decoupling the model into two outcomes - a completed pass ($O_\rho = 1$) or an intercepted pass ($O_\rho = 0$) - will result in the following equation:

$$\begin{aligned} E[X|A = \rho] =& E[X|A = \rho \cap O_\rho = 1] * P(A = \rho \cap O_\rho = 1) && (3) \\ &+E[X|A = \rho \cap O_\rho = 0] * P(A = \rho \cap O_\rho = 0) && (4) \end{aligned}$$

  where equation (3) evaluates the outcome of a completed pass and equation (4) is concerned with an intercepted pass. The expectations in above equation are estimated through machine learning based on spatiotemporal features.

- **Linear Regression**: Pass completion and turnover probabilities are estimated through Linear Regression.

- **Convolutional Neural Network**: The Action-Likelihood model (the probability distribution of action A) is computed through a Convolutional Neural Network.

3.3 Evaluation of EPV

There are several ways that teams can use EPV in order to improve their success. If one player increases his team's EPV most of the time, this is an indicator for this player's decision-making and understanding of the game which can be useful in player signings.

Another idea would be to include analysis of EPV into youth programs of professional soccer teams or in every day practice to teach the players what decisions usually help a team win and which ones not.

4 Evaluating Passes

Passes are the most common events in soccer, thus there is a large sample size to analyze. However, the usual way of recording them is through assigning every pass a binary label (*success* or *no success*) which does not tell the whole story. A pass close to the opponents goal is more difficult to complete than a pass between two defensive players with no opponent nearby.

4.1 Introduction of the paper

Thus, the goal of the paper "Not All Passes Are Created Equal" by Power et al. [PRWL17] is to make an objective analyis to the quality of pass attempts.

It first assigns two parameters to every pass:

- **Risk**: What is the probability that the pass is successful? In mathematic terms, risk is computed as the probability that the pass arrives at your teammate.
- **Reward**: How much would the completed pass help a team score? The reward is a percentage that is the probability that there will be a shot attempt within the next 10 seconds.

Intuitively, backward passes pose a low risk, but they also do not provide a high reward while forward passes could lead to shot attempts but are often intercepted.

4.2 Computing Risk

The authors used several different parameteres to quantify the probability of a completed pass.

- **Naive**: Expecting each pass to be completed with a probability of 85% which is the the average of a completion rate.
- **Using Ball-Information**: Determining the completion probability according to the starting and end point of the pass.
- **Tracking Data**: In addition to ball position, using trajectory of each player and other features that have an impact on the completion. Those factors include obvious things such as:
  - the speed of the player who passes and of the intended receiver
  - the distance of the nearest defender to the passer and receiver

  But it also includes factors that are not too obvious, such as

  - the time from regaining possession: If a team just recovered the ball from the opponent, it is more likely that the opponent is still unorganized, so a completion is more likely.
  - the fact whether the pass is made on the first touch or not. This would decrease the chances of completion because it is technically more difficult.

If a pass is completed, it can be assumed that the receiver was the intended taget. However, if a pass is intercepted by an opponent, there has to be a way to determine who was the most likely target of the pass.

For all players the minimum angle to the ball and the minimum distance to where the ball was intercepted is computed and the intended receiver R^* is determined through

$$R^* = \arg\max_R \frac{d_R}{d_{min}} \times \frac{\alpha_R}{\alpha_{min}} \tag{5}$$

where d_R is the distance of the player to where the ball was intercepted and α_R is the angle between the trajectory of the ball and the line between the passer and receiver R. The index *min* is for the minimum value for all players. This has limitations, for example when a pass is intercepted early in its trajectory or when two players are very close to the ball. Then, the function could not determine the intended receiver correctly but since these scenarios do not happen quite frequently, this is not a very big problem for now. These factors then were classified using a logistic regressor.

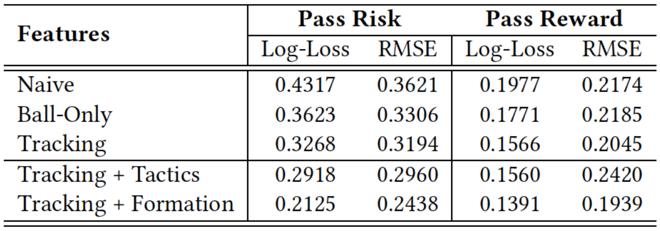
4.3 Computing Reward

As mentioned before, the reward computes the probability of a shot attempt within the next 10 seconds after completion. This window was previously used in different literature [LBM$^+$15] and is used by coaches as a threshold of when a possession should end in a shot attempt.

A negative side effect of this approach is that shots happen very sparsely so very rarely does a shot attempt happen within 10 seconds of a pass. In this experiment, about 5% of the passes are classified with a positive reward.

4.4 Incorporating Context

Knowing more about the context helps the model to make better predictions which can be seen in Figure 3.



Features	Pass Risk		Pass Reward	
	Log-Loss	RMSE	Log-Loss	RMSE
Naive	0.4317	0.3621	0.1977	0.2174
Ball-Only	0.3623	0.3306	0.1771	0.2185
Tracking	0.3268	0.3194	0.1566	0.2045
Tracking + Tactics	0.2918	0.2960	0.1560	0.2420
Tracking + Formation	0.2125	0.2438	0.1391	0.1939

Figure 3: Impact of contextual data

The first three approaches were described earlier and the accuracy increased significantly, especially when computing the risk. However, since soccer is a very tactical sport, more information is needed. One example for that would be the situation or the tactics. Knowing whether the current situation is a

- build-up
- counter-attack
- unstructured play

which is commonly used as distinction in soccer analysis, makes the prediction more precise.

Regarding the last row of Figure 3, the authors determined the tactical formation through the algorithm that will be described in chapter 5.

4.5 Application

There are different applications that the paper describes, one of them being team analysis. This can be done by comparing the quality of passing between teams. For that, the authors used a k-means clustering algorithm to categorize each pass inside or into the opponents half into 16 clusters. Those clusters can be seen in Figure 4.

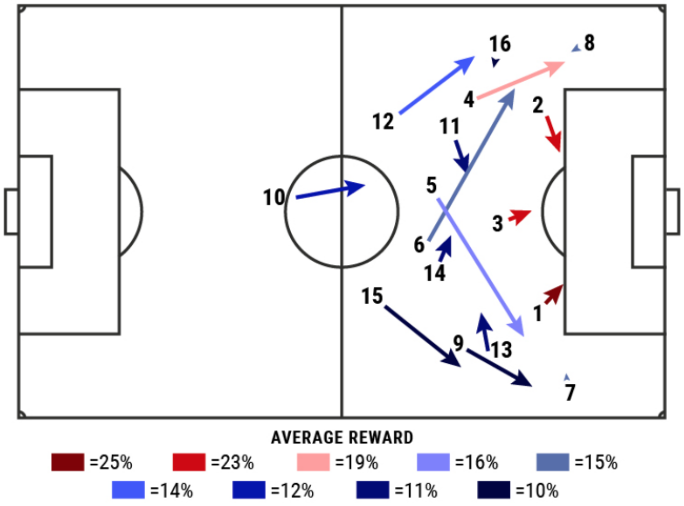


Figure 4: Clustering of passes

Each of those clusters has an average reward ratio, that is visualized through the color of the arrow. Intuitively, passes into the opponents "box" have a relatively large reward rating, though they are also the riskiest ones.

4.6 Evaluation

This model can be used to determine more precisely which players and teams are good at passing and which ones might need improvement. If a manager is looking for players to sign with a small budget, they would find a lot of suitable candidates if they analyzed the players with this method.

When looking at the application described above, a coach can see how well his team is doing in a certain area of the field. If the team's passes in the "dangerous" area of the field are much less successful than average, he might find a reason why his team does not score a lot.

This paper suggests a method with which the event of a pass can be categorized in a highly objective manner and not (as usual) through the eye of a human judge who writes down events when watching the game.

5 Finding Formations

One aim of soccer analysis is to determine the roal of each player on the field depending on their position. In a simpler way that is already possible through so-called "heat-maps" (see left half of Figure 5), where the players' x- ,y-coordinates are plotted on a soccer field. Taking the mean of each player results in the average position. However, this approach is flawed since players often switch positions. If a left-winger switches with a right-winger every few minutes (as it is not uncommon in modern soccer), the average of both players would lie somewhat in the middle of the field. Biatkowski et al. [BLC$^+$14] found a method that assigns a role to the current position of every player. However, the only information needed is the relative positioning of the players with respect to their teammates but not which player it is.

5.1 Introduction of the paper

For this analysis, tracking data from the English Premier League was used. It included the x-, y-coordinates of every single one of the 22 players, recorded at $10Hz$ plus tracking of the ball.

The goal of this paper is to find the most likely formation $\mathcal{F}^*$ given a data set D as in:

$$\mathcal{F}^* = \arg\max_{\mathcal{F}} P(\mathcal{F}|D) \tag{6}$$

5.2 Method

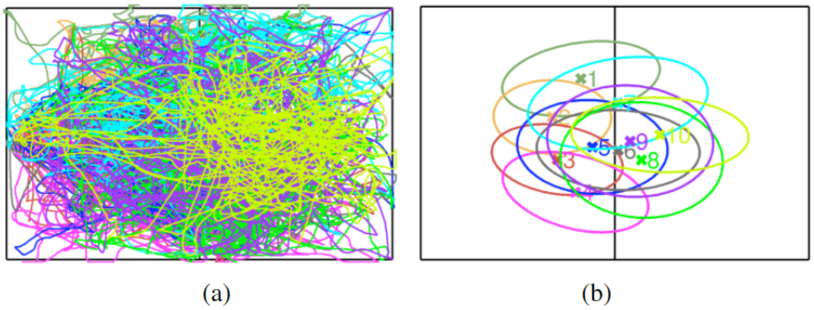
Starting with equation (6), the method uses a concept called *Minimum Entropy Data Partitioning* on the function $P(x)$, which is the probability distribution of the heat-map of the entire team.

$$P(x) = \sum_{n=1}^{N} P(x|n)P(n) = 1/N \sum_{n=1}^{N} P_n(x) \tag{7}$$

where n's are the different roles of the team, and $P_n(x)$ is the heatmap of one role n. Then, it is determined how much the density of the overall heat-map $P(x)$ and the individual heat-maps $P_n(x)$ align. Using the Kullback-Lieber Divergence, which computes the overlap between two probability densities, V_n can be defined as

$$V_n = -KL(P_n(x)||P(x)) \tag{8}$$

$$\mathcal{F}^* = \arg\min_{\mathcal{F}} V \tag{9}$$



(a)

(b)

Figure 5: Using the heatmap (left) to initialize with a rough estimate (right)

When using the entropy on $P(x)$

$$H(x) = -\int_{-\infty}^{\infty} P(x) \log P(x) dx \tag{10}$$

the optimization problem can be rewritten after a few substitutions as

$$\mathcal{F}^* = \arg\max_{\mathcal{F}} \sum_{n=1}^{N} H(x|n) \tag{11}$$

The goal then is to find 2D-Gaussian distributions for the 10 outfield players (ignoring the goalkeeper) that represent the role of each position. It is initialized through the 2D-Gaussians of each player. As stated before, the roles should only be attributed to a position on the field not a player, since players regularly switch positions or get substituted for another player. As an initialization (see Figure 5), this seems reasonable, anyway.

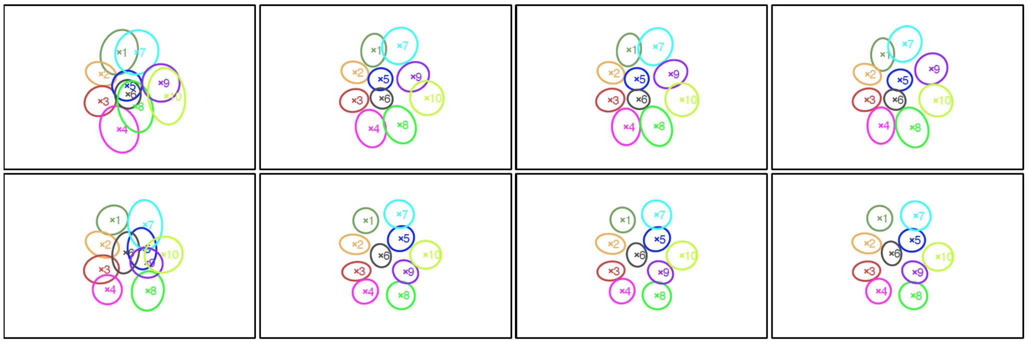
The method then uses an algorithm similar to k-means clustering, with the only difference that at every frame, each location is assigned a role in the formation.

In each frame t of the tracking data every pair of (x_t, y_t) is assigned a role. This is done through a cost matrix M based on the probability that a current position has a certain role label. This process is repeated as an expectation maximization algorithm (EM) until the formation has converged. The final formation then consists of 10 distinct role probability distributions.

5.3 Results

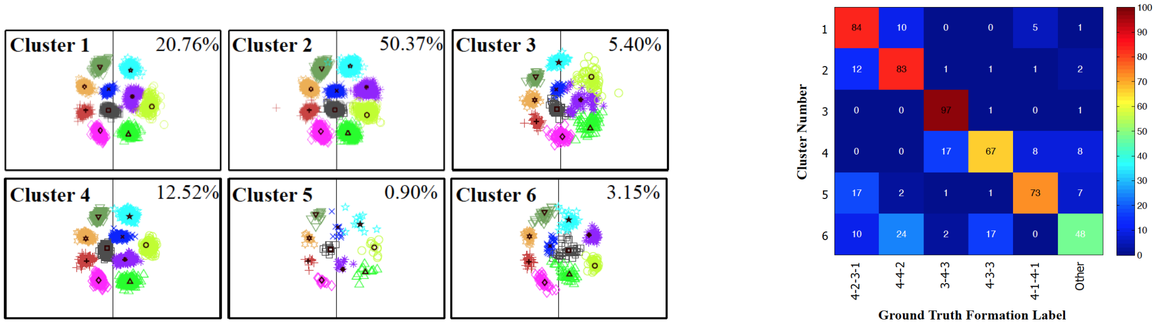
The development and convergence of the algorithm can be seen in Figure 6.

Evidently, the formations converge more and more with every iteration until a clearly visible formation can be detected in the final iteration.



(a) Initial Roles (b) Iteration 1 (c) Iteration 2 ... (d) Final Roles

Figure 6: Convergence of EM-algorithm



(a) Clusters of the positions (b) Error-testing

Figure 7: Clustering the formations and error-testing

Now, it is interesting to see what formations were used across the league and how often each formation is used. This can be categorized into 6 clusters - the 5 most common formations and the ’other’ category - each having a similar shape, see Figure 7a.

In order to compute the correctness of the algorithm, a soccer expert was given the task to determine the formation in each situation. These results were then compared to the results of the algorithm. It can be seen in Figure 7b that the algorithm got most formations correctly, often above 83%.

5.4 Evaluation

This paper attempted to make a role-based analysis unlike a position-based analysis as usual in the past. The key aspect in this analysis is the minimization of the entropy of positions. This is prone to small changes during a match in which players switch positions or move somewhere else on the field. This opens the door for a wide range of possibilities, such as analysis of team tactics or individual tactics.

6 Conclusion

Unlike other sports such as Baseball where events happen in a discrete order, soccer is a very fluent sport that is very difficult to analyze mathematically. However, several researchers have attempted to design models that can give huge insight into how the sports work. One of the key components in this kind of analysis is tracking data.

Using mathematical models that take the subjectivity of a human judge away, can give objective results that - if used correctly - teams on a low budget could use in order to find new players that are being overlooked. For example, a player called Wesley Hoolahan was one of the most efficient passers in the English Premier League and yet his transfer value never exceeded 3 million Euros [TM119].

Professional clubs could implement strategies in their youth programs to teach young players about which actions such as passes lead to a higher success rate than others. Teams can also prepare for their next opponent by analyzing what type of passes they like to play (more risky, more backwards etc.).

There is a lot of potential in soccer analytics, which hasn't arrived completely in the minds of sports executives but it can give a team a tremendous competitive advantage.

References

[BLC+14] Alina Bialkowski, Patrick Lucey, Peter Carr, Yisong Yue, Sridha Sridharan, and Iain Matthews. Large-scale analysis of soccer matches using spatiotemporal tracking data. *International Conference on Data Mining (ICDM)*, December 2014.

[FBC19] Javier Fernandez, Luke Bornn, and Dan Cervone. Decomposing the Immeasurable Sport: A deep learning expected possession value framework for soccer. *MIT Sloan Sports Analytics Conference*, 2019.

[LBM+15] Patrick Lucey, Alina Bialkowski, Mathew Monfort, Peter Carr, and Iain Matthews. "quality vs quantity": Improved shot prediction in soccer using strategic features from spatiotemporal data. 2015.

[Lew04] Michael Lewis. *Moneyball: the art of winning an unfair game.* W.W. Norton, 2004.

[PRWL17] Paul Power, Hector Ruiz, Xinyu Wei, and Patrick Lucey. "not all passes are created equal:" objectively measuring the risk and reward of passes in soccer from tracking data. *Proceedings of KDD conference, El Halifax, Nova Scotia Canada*, February 2017.

[TM119] https://www.transfermarkt.com/wesley-hoolahan/profil/spieler/24589, December 2019.