uc3m | Universidad Carlos III de Madrid

MASTER IN STATISTICS FOR DATA SCIENCE

2023-2024

Master Thesis

Introducing Offside Control: a Football Analytics Parameter to Evaluate Offside Performance

Álvaro Novillo Correas

Juan Miguel Marin Diazaraque

Javier Martín Buldú

Madrid, June 9, 2024

CONTENTS

ABSTRACT

This master thesis explores the application of data analytics in football, with a focus on enhancing understanding and performance through a data driven approach. The study reviews the history and evolution of data analytics in sports, particularly football, and introduces a new methodology for quantifying the effectiveness of offside strategies using Pitch Control models, *Offside Control*. We analyze 1,251,934 frames from 100 matches to characterize the Offside Control of 442 players. The methodology combines physical metrics and tactical models derived from tracking datasets, offering insights into the tactical nuances of successful offside strategies. The project's data processing and modeling are conducted using Python, with the source code available in a GitHub repository[1].

[1] All the source code is inside the code folder at https://github.com/AlvaroNovillo/master_thesis.git

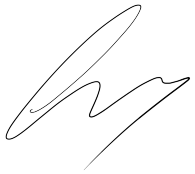
DEDICATION

To Prof. J.M. Buldú, whose trust and belief in my abilities opened doors I could never have imagined.

To my dearest friends and family, for always trusting and believing in me much more than I do.

And to Prof. J.M. Marín, whose dedication and guidance illuminated the path of this thesis. The thesis felt lighter under the laughs we had in the way.

With sincere gratitude,



Álvaro Novillo

Madrid, May 2024.

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

The digital revolution is currently one of the most significant challenges of our time, altering numerous aspects of society. Sports, in particular, has also been influenced by this transformation. Technological advancements and digitalization have resulted in a swift upsurge in the number of measuring devices, data collection and volumes of data. The leading data companies worldwide, including IBM, Intel, SAP and Microsoft, are vying for superior data analytics tools and leveraging sports as an example domain to showcase their products and brand power (Footballytics, 2021).

Basketball is one of the best examples of data analytics being applied to sports. Teams use data to analyze player performance, identify strengths and weaknesses, and develop strategies to win games (Sarlis & Tjortjis, 2020). Basketball Analytics can provide qualitative analysis to team owners, players, coaches, and technical staff to help them predict future situations and make informed decisions to improve performance. Such analysis have become vital to a team's success, as it aims to reduce expenditure, enhance team worth, and refine processes across all levels and segments of operations. The German Football Association (DFB) and the National Basketball Association (NBA) are two examples of digital transformation in the sports world. Successful teams use player performance data to gain a competitive advantage.

In the case of football, the practice of data analytics has a long history, dating back to the post-World War II era, when data collection and analysis was undertaken manually using pencil and paper (Footballytics, 2021). It was not until Moneyball was published in 2003 that significant progress began to emerge: The book, "The Art of Winning an Unfair Game" introduced sports analytics to a broader audience. It illustrated the use of data analytics in identifying undervalued players and constructing a successful team. Since then, data analytics has become an integral component of sport (Footballytics, 2021).

Over the last years, football analytics has gained significant popularity, aiming to delve deeper into the game by utilizing advanced data analysis techniques to optimize team and player performance.

The main objective of this master's thesis is to enhance understanding and performance in football through the use of data analytics. The master's thesis includes a literature review of the field, alongside the commonly found data types within this industry and the main metrics used to analyse player and team performance, focusing on tracking data and Pitch Control models (Spearman, 2018). Following the initial review of the field's state of the art, we propose a new methodology for quantifying the effectiveness of the offside strategy of teams and players using Pitch Control models. Our study defines a new performance parameter, called *Offside Control*, which quantifies the amount of threat posed by the attacking team or player beyond the offside line.

In the study, we will compute both effective and ineffective Offside Control at a rate of 2 frames per second for 100 matches from LaLiga 2019-2020, resulting in a total of 1,251,934 frames analysed. This will allow us to characterize successfully the Offside Control of 442 players in total.

Our proposed methodology aims to contribute to this growing body of knowledge. Analyzing vast amounts of tracking data from LaLiga matches, we seek to uncover patterns in player and team behavior, shedding light on the tactical nuances that underlie successful offside strategies.

We presented this metric at *OptaForum 2023*, which took place in central London on Tuesday 21st of March. Our proposal was one of only 5 selected to be included in the congress. It was a privilege to learn from other leading experts in the field, and also generated fruitful conversations that were fundamental to the development of the work[2].

All data processing and modeling in this project has been made using python. You can find the source code

[2]You can check the rest of the proposals accepted here https://www.statsperform.com/resource/first-presentations-announced-for-the-2023-opta-forums-pro-track/

of the project inside the following GitHub repository: https://github.com/AlvaroNovillo/master_thesis.git

2. DATA ANALYTICS IN FOOTBALL

When discussing sports analytics in football, the first metric that often comes to mind is the *Expected Goals (xG) ratio*. This statistical indicator is a predictive Machine Learning (ML) model used to assess the likelihood of scoring for every shot made in the game. In the context of each shot, most Expected Goals models consider the following factors: distance to goal, angle to goal, body part used for the shot, and type of assist or previous action (such as throughball, cross, set-piece, or dribble). Then, using a regression model, the xG model assigns a value between 0 and 1 to each shot, indicating the probability of it resulting in a goal.

Wyscout xG model, for example, encompass the shot's spatial coordinates, the assisting player's position, the striking player's use of foot or head, the type of assist involved, the occurrence of a dribble by either a field player or the goalkeeper immediately preceding the shot, whether the shot arises from a set piece, whether it transpires during a counterattack or in a transitional phase of play, and the subjective assessment of shot danger as determined by a designated tagger. This set of parameters serves as the foundation for training the xG model using historical Wyscout data, culminating in the prediction of the likelihood of a given shot resulting in a goal (Wyscout, 2023).

Since the probabilities range from 0 to 1, a shot with an xG value of 0.1 has a 10% chance of being scored. Penalties have a fixed xG value of 0.76.

Fig. 2.1 provides a visual representation of the cumulative development of expected goals (xG) during the Eibar - Malaga match, which took place on January 15th, 2023 in Spain's second division. Each data point on the graph corresponds to a shot made by both teams during the game, offering a comprehensive overview of the evolving scoring opportunities and outcomes throughout the duration of the game.

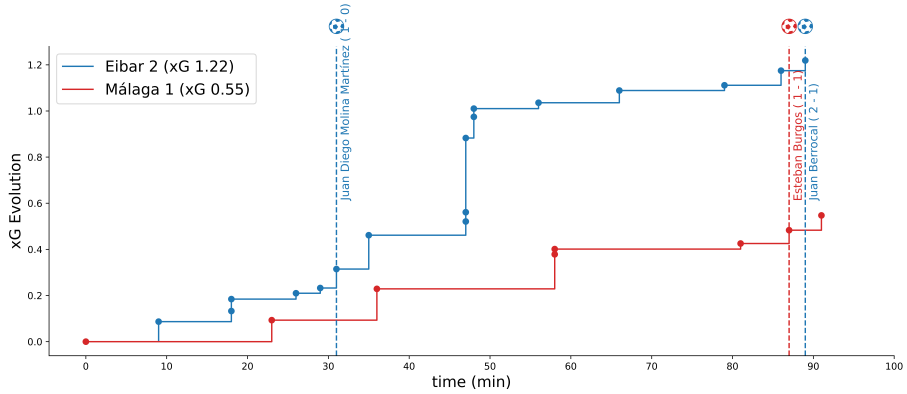


Fig. 2.1. Cumulative development of expected goals (xG) during the Eibar-Malaga match, held on January 15th in Spain's second division. Each point denotes a shot made by both teams throughout the game. Vertical dashed lines indicate the goal scored, displaying the player and the corresponding score at that specific moment of the match.

Expected Goals (xG) have revolutionized the analysis of football by quantifying the quality of scoring opportunities. However, it is important to consider other variables such as player positioning, velocity, passing accuracy, defensive pressure, and tactical formations to gain a broader understanding of the sport.

Looking at an xG evolution figure, such as Fig. 2.1, and solely focusing on shot probabilities while disregarding the spatial distribution of shots and occasions feels like merely scratching the surface of what sport analytics can offer to football.

To illustrate the spatial distribution of shot locations taken by both teams during the game, we can create a shot map for each shot. In Fig. 2.2, the size of each data point corresponds to the expected goals (xG)

generated for the respective shots, providing insights into the perceived scoring potential. Goals scored are visually highlighted with a ball marker, and straight lines indicating the trajectory the ball followed as it found its way into the opponent's net. Bellow each shot map, a plot of the net can be also found, where goals are represented by football balls, and blocked shots by shadowed points. This detailed analysis not only enhances our understanding of scoring opportunities but also sheds light on the tactical strategies employed by both teams, player positioning, and defensive vulnerabilities. Analyses such as the one above are carried out using the most common source of data in football: **Events** datasets.

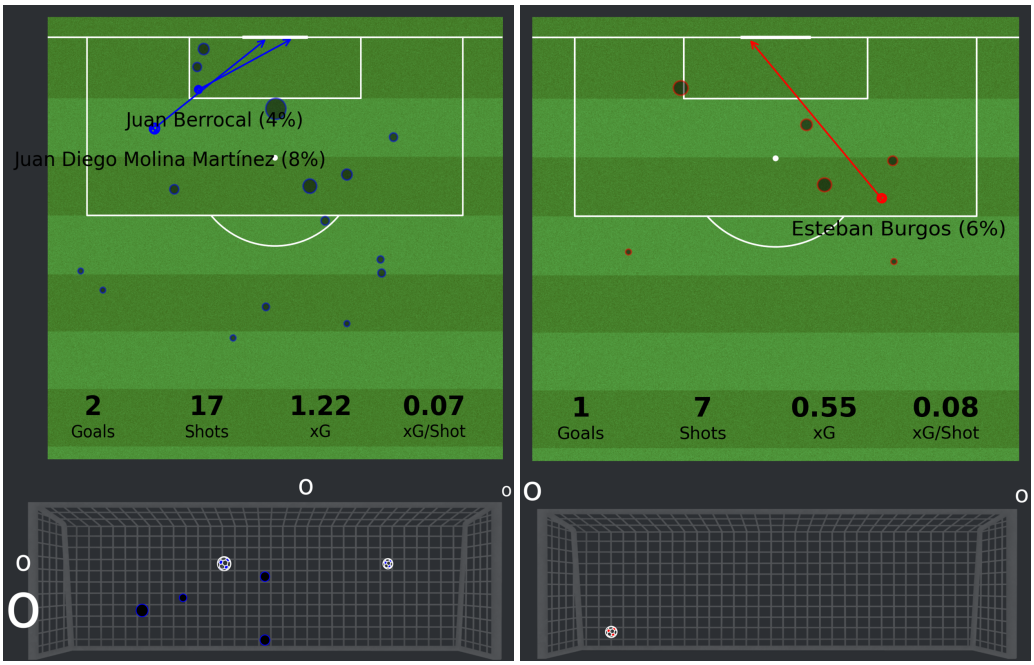


Fig. 2.2. Shot map of the Eibar (blue, left) - Malaga (red, right) football match. The locations of the points indicate where shots were taken. The size of each point is proportional to the expected goals (xG) generated. Shots that resulted in goals are depicted with a straight line, representing the path the ball took to enter the opponent's net.

2.1. Events data

`Event data` describes specific, human-defined events during a match, including passes, shots, and fouls. It is captured by human annotators from various providers. However, this manual process is time-consuming and typically requires three individuals:

The data collection process is carried out by professional video analysts (known as operators), who are specialists in football data collection, using proprietary software (the tagger)[3]. The tagger has undergone several years of development and improvement and is regularly updated to ensure the highest level of performance is achieved. To ensure accurate data collection when tagging events in football matches, three operators are assigned: one per team and one supervising the output of the entire match. This process is based on analysis of the tagger and football match videos. When near-live data delivery is necessary, a team of four operators may be utilized, with one operator dedicated to hastening the collection of complex events that require additional, specific attributes or a quick review (Pappalardo et al., 2019)

Table 2.1 displays some rows of an event record for a game. Each event is tagged with a code, and the (x, y) coordinates of the starting position of the event are given, as well as the time at which it occurred and the player who performed the event.

This type of data structure can be used in a number of ways: it can be used to measure team performance through general statistics extracted from event datasets, such as goals, fouls, xG, etc. It can also be used to

[3] See Fig. 1 in (Pappalardo et al., 2019) for an example of the tagging software

team_id	code	type	player	x	y	min	sec	period
174	44	0aerial_duel_lost	18498	71.2	25.5	0	7	1
174	5	0ball_out	18498	77.0	-1.7	0	12	1
178	5	0ball_out	171101	70.0	-1.2	0	45	1
178	12	0despeje_clearance	52356	29.5	85.1	1	6	1
178	5	0ball_out	52356	25.2	101.3	1	7	1
174	44	0aerial_duel_lost	18498	64.2	42.4	1	27	1

Table 2.1. EXAMPLE OF AN EVENTS DATASET OF A GAME.

create advanced analysis of the team using ensembles of mathematical models.

The match analysis is frequently conducted using Graph Theory, (Buldú et al., 2018), (Novillo et al., 2024). Combining different elements of the events dataset, we can create a graph corresponding to the passing network of each team, allowing us to understand the passing structure of both teams.

Figs. 2.3 illustrate the passing networks observed in the Eibar versus Málaga football match, providing insight into the passing interactions and tactical strategies used by both teams. The nodes in the graphs represent individual players who participated in the match for each team. The nodes are sized according to their degree, which represents the amount of ingoing and outgoing passes. The node position corresponds to the average passing position of each player. Substitute players are represented by yellow nodes, and links are created if there have been at least 5 passes made in that direction between two players. The edge's width corresponds to the amount of passes made in that direction between the two players.

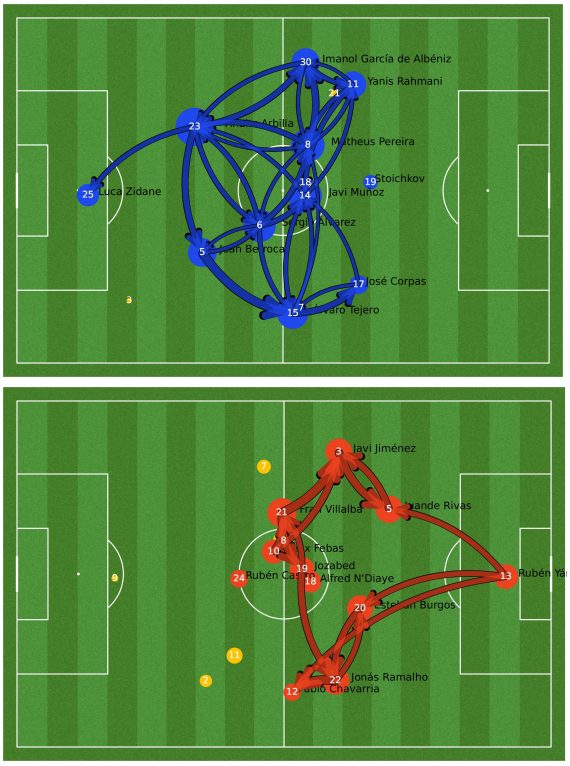


Fig. 2.3. Representation of the Eibar (blue, left), Málaga (right, red) passing networks of the match Eibar - Málaga. Nodes represent players, edges represent passes between players. The position of the players in the field is their average passing position. The size of the nodes reflects the number of ingoing and outgoing passes (i.e. node's degree), while the size of the edges is proportional to the number of passes between the players. Substitute players are represented in yellow. A connection is set if those players share at least 5 passes. The edge's width is proportional to the amount of passes made in that direction between the two players.

Analysis as the former can be conducted *in real-time*[4] during the match using appropriate data sources. Additionally, we could examine Eibar's macro situation during the 2022-2023 season to better comprehend how this micro-statistics contribute to the overall perception of the team.

Fig. 2.4 presents the expected goals (xG) produced by Eibar in all matches played against their opponents. It is noticeable that Eibar has generated a higher xG when playing at their Home stadium, on average.

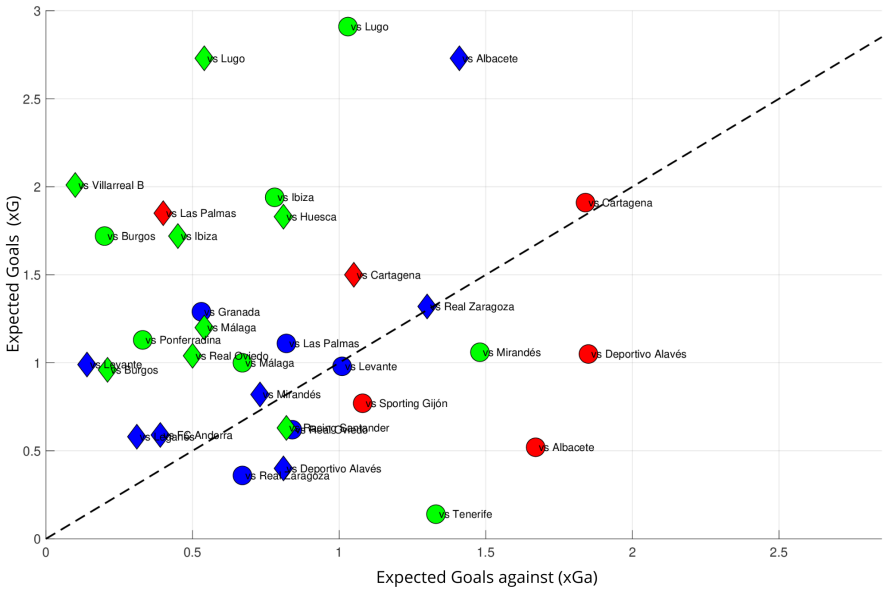


Fig. 2.4. Expected Goals (xG) and Expected Goals Against (xGA) per match. Codes: Home Matches (Diamonds), Away Matches (Circles), Wins (Green), Draws (Blue), Losses (Red). Matches above the dashed lines represent those matches where Eibar has generated more xG than the opponent.

So far, we have discussed some of the statistics that can be inferred from this data source to characterize team performance, such as possession, pressure, duels, and fouls. These datasets are easily manageable and provide teams with useful information. However, due to limitations in space and scope, we cannot provide a more in-depth analysis of these measures.

Although event datasets supply beneficial information regarding the team's overall performance, deeper scrutiny can be conducted via tracking data, which consists of the players' and ball's position and movement during the match. Tracking data can offer additional insight into both the physical and tactical aspects of the game.

2.2. Tracking data

`Tracking data` offers a more comprehensive perspective than event data by providing access to information on all players, their trajectories, and velocities. This allows for the analysis of off-ball players and team dynamics, resulting in a more nuanced understanding of the game.

There are two main techniques for obtaining tracking data, which decide its classification: **Image detection algorithms** extract players' positions from the match broadcast and infer locations of players that are not visible in the broadcast frame, whereas **optical tracking** employs a specialized camera system installed on the field to record players' data. Our research will concentrate on the latter method, as it offers more precise and statistically informative data.

[4]Opta (StatsPerform, 2023) uses a combination of human annotation, computer vision, and AI modelling to offer real-time data at various levels of detail based on customer requirements. In our situation, the data feed updates itself when an event such as a goal, foul or pass occurs. Otherwise, it updates every 90 seconds.

Our tracking data has been provided by Mediacoach®. They utilises the Tracab Optical Tracking system to obtain on-the-pitch player positions. This multi-camera system captures each player's position at 25 frames per second. The system consists of three units, each with a resolution of 1920x1080 pixels, producing a panoramic picture that generates a stereoscopic view for triangulating the players and ball. In case of a temporary loss of any location, a skilled operator adjusts the players' positions. The datasets obtained by the Mediacoach® system have been validated in advance using GPS (Felipe et al., 2019).

Fig. 2.5 contains a frame from the tracking dataset of an Atlético de Madrid (Blue) - Getafe (Red) game from the Spanish 2019 League. The ball is shown as a black dot. Referees are shown as yellow squares. Purple arrows represent the speed vectors of the players.

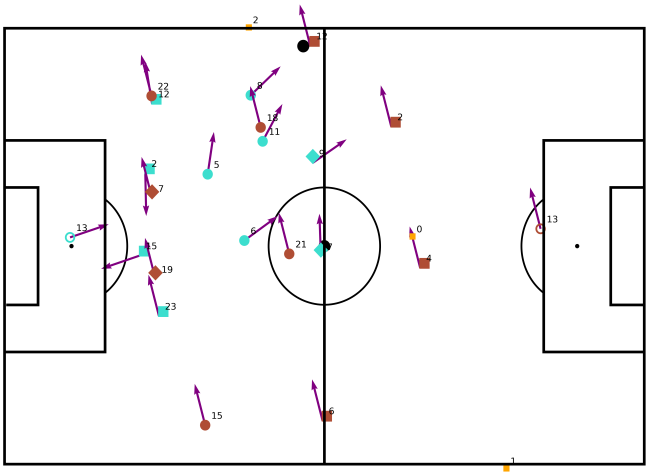


Fig. 2.5. A frame of tracking data from a football match. The home team is shown in blue, the away team in red. The ball is shown as a black dot. Referees are shown as yellow squares. Purple arrows represent the speed vectors of the players.

Traditionally, all football statistics have been produced using event datasets. In this respect, tracking datasets have been crucial in developing new ways of measuring team and player performance. With this type of information, we are not just limited to ball-related events such as passes, goals, etc. tracking datasets contain the precise location of players and the ball during a match. (Garrido et al., 2022) showed that the correlation between heatmaps, Fig. 2.6, made with event datasets and those made with tracking datasets is low. These heatmaps are constructed by discretising the field using a grid and measuring how frequently the object at study has been in such discretized spaces. Event heatmaps show where a player has performed more actions. On the other hand, tracking heatmaps can provide information about a player's position without adding information about how relevant their performance was, since we have their position at every instance of the game.

2.6 uses tracking data to show the position of the ball during the match between Atlético de Madrid and Getafe. With a simple glance, we can extract a lot of information from the match. The ball was most likely to be found in the centre-left part of the field, showing a strong attacking dominance from Getafe. Note that we always keep the direction of play from left to right, so the home team will always be placed on the left of the pitch and the away team on the right.

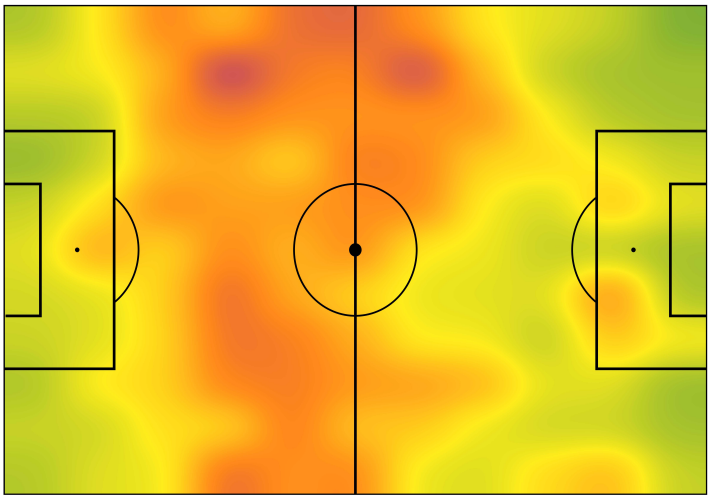


Fig. 2.6. Heatmap of the ball position during the Atlético de Madrid - Getafe game under study. Note that we always keep the direction of play from left to right, so the home team will always be placed on the left side of the field and the away team on the right.

Thus, tracking datasets are crucial for developing new ways of measuring team and player performance, as they provide us with global information about the game.

This data-driven approach enhances the understanding of the sport and its strategic nuances, fostering a deeper appreciation of the game's intricacies. As a result, tracking data has become a valuable tool in football analysis, providing insights into player performance and team strategy. It can be applied to specific games, as in the analysis presented earlier (Fig. 2.6), or to an ensemble of them to provide a comprehensive view of general player behavior under different physical and tactical metrics, constructed using tracking datasets.

In the next chapter, after this initial review of the state of art of football analytics, we will combine both physical metrics and tactical models derived from Tracking datasets to build our proposed metric to characterize offside strategy's effectiveness: *Offside Control*

3. METHODOLOGY

As discussed earlier, tracking datasets provide the position of all players and the ball throughout the match with a temporal resolution of 25fps. This enables us to estimate the players' covered distance, speed, and acceleration. The potential for extracting information from tracking datasets that is useful for football analytics extends beyond variables related to players' physical performance. Many tactical metrics has been implemented to decode how the intricate movements of players translate to the football field.

These models provide a scientific perspective for analysing player positioning, decision- making, and team dynamics, illuminating the complex interactions that occur during a match.

Pitch Control is one of the most relevant tactical metrics used to analyse player positioning, decision-making, and team dynamics during a football match. It combines player position and speed with mathematical models that simulate ball and player movement (Spearman, 2018).

This thesis proposes to use Pitch Control Models to evaluate how soccer teams interact with the offside line when attacking and defending. Before doing so, we first need to define Pitch Control Models, including their construction and implementation.

3.1. Pitch Control Models

The *Pitch Control (PC)* at a given location represents the probability of a player or team gaining control of the ball if it moves directly to that location. *PC* models simulate the dynamics of the ball and the players to evaluate which player would control the ball if it moves to any location on the pitch at any moment. The model captures not only the players' current position, but also their movement. When players are running at high speeds, they are more likely to control the space they are moving into rather than the space they currently occupy.

To construct this model, we must address the following issues for a given location on the pitch:

- How long it would take for the ball to reach to the position of interest (from its starting position).
- How long would it take for each player to get to that position.
- What is the total *probability* that each team will control the ball *after* both the players and the ball have arrived at the desired position?

In the model, the ball is set to move at a constant speed of $v_b = 54\ km/h$, which is approximately the average speed of the ball in the game (Spearman, 2018) (See Fig. 3.1)

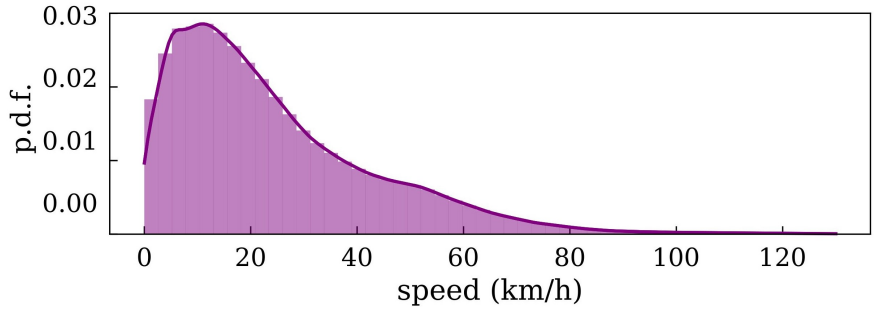


Fig. 3.1. p.d.f of the ball speed over a 100 matches from LaLiga 2019-2020 season.

Therefore, the time taken to arrive at the location of interest can be easily calculated as $t_{b,arr} = \Delta x_b / v_b$, where Δx_b is the distance between the initial and final positions of the ball.

3.1.1. Model assumptions

When considering how long it will take the players to reach the target position, given their initial position and speed, players are assumed to only have a maximum speed of $v_{max,p} = 18\ km/h$, which corresponds to the 95 percentile of the average speed of the players in the game[5] (Spearman, 2018) (See Fig. 3.2). This upper limit should not be misunderstood as the maximum speed at which players can move, but rather as an estimate of the maximum speed at which they are likely to move when trying to control the ball

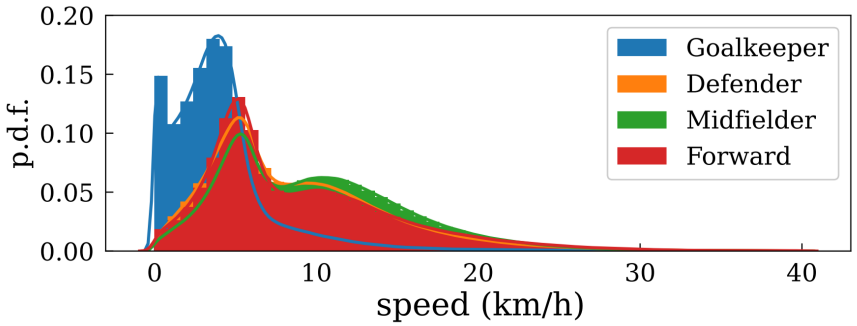


Fig. 3.2. p.d.f of the players speed over a 100 matches from LaLiga 2019-2020 season.

To compute the player's expected arrival time, $\tau_{exp}(\vec{r}; t_r)$, we use a simple approximation consisting of a two-step process:

- There is an initial *reaction time*, assumed to be of $t_r = 0.7$ seconds for every player[6]. This is approximately the time it takes a player moving at maximum speed to come to a complete stop (Spearman, 2018). During this reaction time, we assume that players continue to move along their current trajectory without changing speed or direction (reaching a position $\vec{r}_{\text{react}}$).
- After this time, we assume that the player runs directly towards the ball at his maximum speed of $v_{\text{max,p}}$.

$$\tau_{exp}(\vec{r}; t_r) = t_r + \frac{|\vec{r} - \vec{r}_{react}|}{v_{max,p}}$$

3.1.2. Control probability

Once we have computed the time it takes for the ball and the players to get to the target location, we need to look at how long it will take each player to control the ball. To do this, we assume that controlling the ball is a stochastic process that follows an exponential distribution with a fixed rate λ. The inverse of such a free parameter, λ^{-1}, can be thought of as the time it takes a player to gain control of the ball, in seconds. Modeling the process with a exponential distribution, we capture the fact that players who stay closer to the ball for longer are more able to control the ball, (3.1) (Spearman et al., 2017). Thus, for any differential time Δt that a player is near the ball, he has a probability of $\lambda \cdot \Delta t$ of controlling the ball.

$$F(\Delta t; \lambda) = 1 - e^{-\lambda \Delta t} \tag{3.1}$$

So far, the model assumes that we know exactly when each player will arrive at the target location. However, we introduce some uncertainty, labelled σ, in the arrival time of the players. The reason for including such

[5]We can impose this assumption for all the players without loss of generality, to simplify the model. Further work to improve the model will be individualize this reaction parameter for each player.

[6]We can impose this assumption for all the players without loss of generality, to simplify the model. Further work to improve the model will be individualize this reaction parameter for each player.

temporal variability in our model is to account for some effects that have not been explicitly modeled, such as player effort. To model that uncertainty, we will use a Logistic distribution. Thus, the probability of a player intercepting the ball at time T is given by the cumulative distribution function of the sigmoid distribution (Spearman et al., 2017) (See Fig. 3.3).

$$F_{\text{int}}(\vec{r}, T; \sigma, t_r) = \frac{1}{1 + e^{-\frac{T - \tau_{exp}(\vec{r}; t_r)}{\sqrt{3}\sigma/\pi}}}$$

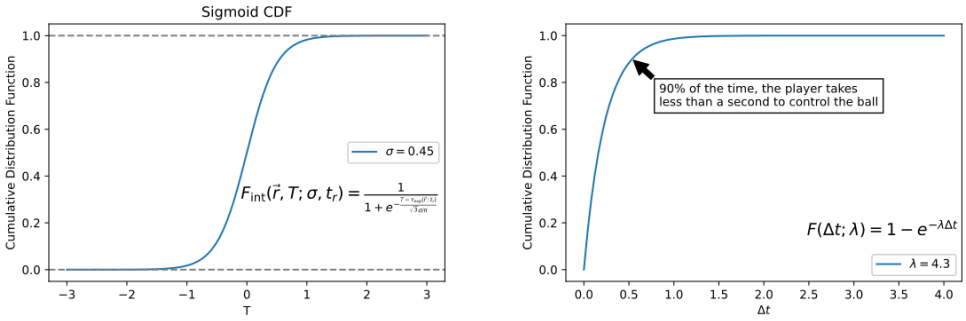


Fig. 3.3. The cumulative distribution functions for the two components of the model. a) (left) the time to intercept and b) (right) the time to control the ball. The parameters shown for each are from the global fit described below.

Both λ and σ has been selected according to (Spearman et al., 2017), where they model passes as a Bernoulli trial, with probability mass function

$$P(k \mid \sigma, \lambda, x) = \begin{cases} 1 - p \text{ for k} = 0 \\ p \qquad\qquad\qquad \text{for k} = 1 \end{cases}$$

where $k \in [0, 1]$ is the outcome of the pass. Then, the likelihood of a set of parameters, σ and λ, given outcome k and the start of the pass x is:

$$\mathcal{L}(\sigma, \lambda \mid k, x) = P(k \mid \sigma, \lambda, x)$$

Then, maximizing the product of the likelihood for each pass for a training sample P from tracking and event data from the 2015-2016 Premier League season, the best fit is found at[7] $\sigma = 0.45 \pm 0.05$ s and $\lambda = 4.30 \pm 1.14$ s^{-1}

$$\min_{\sigma,\lambda \in \{\mathbb{R},\mathbb{R}\}} \left\{ -\sum_{i \in P} \log\left[\mathcal{L}(\sigma, \lambda \mid k_i, x_i)\right] \right\}$$

Using the above components, we recursively construct the partial derivative of the probability that player j controls a given location r, at time t is

$$\frac{dPPCF_j}{dt}\left(T, \vec{r}, \sigma, \lambda_j, t_r\right) = \left(1 - \sum_k PPCF_k\left(t, \vec{r}, \sigma, \lambda_k\right)\right) F_{int,j}(t, \vec{r}, \sigma, t_r)\lambda_j \tag{3.2}$$

[7] See (Spearman et al., 2017) for further details

where $PPCF_j$ is the Potential Pitch Control Field of player j. $F_{int,j}(t, \vec{r}, T; \sigma, t_r)$ is the probability that player j can reach the target location r in a given time t, and λ_j is the control rate of such a player [8]. Importantly, note that $\sum_k PPCF_k\,(T, \vec{r}, \sigma, \lambda_k)$ accounts for the sum of the Potential Pitch Control Field of the rest of the k players on the pitch at time t.

By integrating the equation above, Eq. (3.2), $t \in [t_{ball}, t_{ball} + 10]$ seconds, and taking $PPCF_j\left(t, \vec{r}, \sigma, \lambda_j\right) = 0$ at the beginning of the integration, the probability of control per player is obtained. This probability is then extracted along all the pitch, obtaining a pitch control surface.

Now that we are able to generate the pitch control surface for a frame, we can extend this methodology to measure the quality of each team's offside strategy. To do this, we will focus on the pitch control generated *after* the offside line, called **Offside Control (OC)**. To do this, we determine where the offside line is and which players are in an offside position. Then we calculate the pitch control generated by the attacking team after the offside line. If a player is in an offside position, we mark his contribution as ineffective.

The *Offside Control* was calculated for 100 matches from LaLiga (season 2018 - 2019) using tracking data. We calculated the OC every 2 frames per second of each match. In order to discretise the space efficiently, we decided to reduce our calculations to the half of the pitch of the team not in possession of the ball. We then detect the position of the offside line and calculate the pitch control after it. In this way, we significantly reduce the computation time of the model while maintaining a high spatial resolution (50×32).

[8] We assign the goalkeepers to have a higher control rate, $\lambda_{GK} = 12.9\ s^{-1}$, to ensure that they are likely to claim the ball if it is near them and also to account for the ability of grabbing the ball with their hands

4. RESULTS

The procedure for constructing the *Offside Control (OC)* parameter involves first identifying the offside line, which is determined by the position of the second-last defender. Then we identify the attacking players who are within the allowed boundaries and in an offside position in each frame. Once properly identified, we compute the OC metric as a measure of the field control exerted by the attacking team or player after the offside demarcation.

Additionally, based on each player's location relative to the offside line, we can evaluate a player's involvement in an offside position as a manifestation of ineffective offside control (IOC). Conversely, instances where a player is positioned beyond the offside line and maintains control of the ball are classified as effective offside control (EOC).

4.1. Offside Control

The concept of effective and ineffective Offside Control is illustrated in Fig. 4.1.

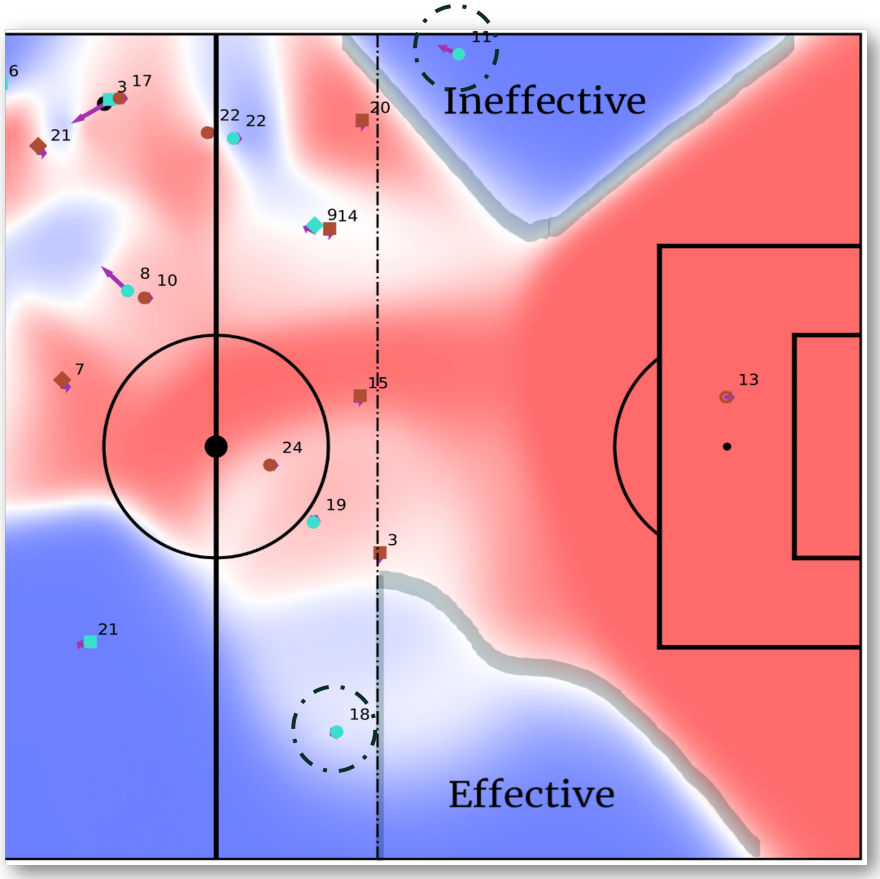


Fig. 4.1. Example of effective (EOC) and ineffective (IOC) Offside Control (areas surrounded by a thick grey line). Offside is indicated by the vertical dashed line. Note that player 11 is offside (generating IOC), while player 18 is not (generating EOC).

Players from both teams are shown in blue (home team, attacking from left to right) and red (away team, attacking from right to left). Areas of the pitch controlled by the home team are shown in blue, while areas controlled by the away team in that particular match are shown in red. In this frame, the home team is in possession of the ball.

A vertical dashed line shows the offside line's location, and the areas enclosed by a broad grey line represent zones controlled by the attacking team behind this line.

Here, the home player wearing number 11 is behind the offside line in an invalid position, thereby generating IOC around him. In contrast, player 18 is correctly positioned, leading to effective Offside Control (EOC) behind the offside line.

4.2. Measurement of Offside Control

The *Spatial Offside Control* generated by a team can be measured by simply adding up the Offside Control values of each team during the match within the regions behind the offside line.

We have also computed the *Percentage of Offside Control* accumulated by each team to measure which team achieved greater control over the opponent.

Fig. 4.2 shows two examples of both spatial OC and percentage of OC for two different games (A and B) by a team called α.

A heatmap highlights in red the spatial OC generated and received by team α, pinpointing specific pitch areas controlled behind the offside line.

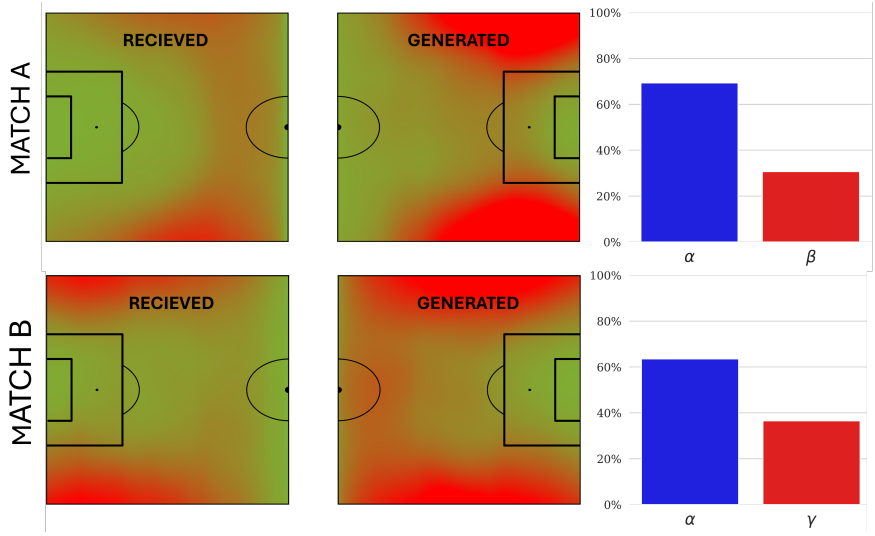


Fig. 4.2. Spatial and percentage of effective offside control (EOC) over the course of two matches (A and B). The left plots show the EOC received by team α and the middle plots show the EOC generated by team α. In all plots, the intensity of the red color is proportional to the accumulated EOC at each location on the field. The bars on the right show the percentage of EOC accumulated by team α and its opponents.

Note that only effective Offside Control is shown in this example. We can see that in match *A* the danger generated behind the offside line is deeper and closer to the box, while the EOC received is not as close to the own goal. On the other hand, in game *B*, the EOC generated is more longitudinal and closer to the edge of the pitch, while the EOC generated by the opposing team is much deeper and therefore more dangerous. Beyond the spatial location, the bars on the right side of the figure 4.2 show the proportion of EOC accumulated by team α and its opponents in both games. In both cases, team α dominates the opposing team, accumulating a percentage higher than 65%.

4.3. Dynamic Monitoring of Offside Control

Offside Control can also be monitored throughout a match. Fig. 4.3 shows an example of $OC(t)$ for the home team of a match, smoothed with a centered sliding window of 5 minutes.

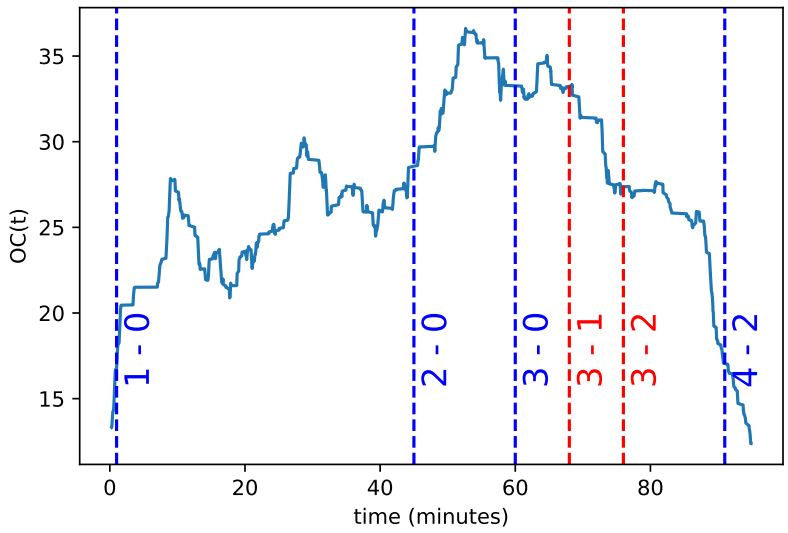


Fig. 4.3. Analysis of the home team's Offside Control over the course of a match. Each point of $OC(t)$ is calculated using a centred sliding window of 5 minutes. The vertical dashed lines indicate the moments when a goal was scored (blue for the home team and red for the away team).

The figure also marks moments when goals were scored, linking $OC(t)$ to the game's score.

As we have plotted the home team's $OC(t)$, we can see that it increases until a few minutes before the third goal is scored. However, it drops around minute 50, which shows that the home team is creating less danger behind the opponent's offside line, presumably trying to defend the goal advantage rather than create new scoring opportunities.

4.4. Comprehensive Analysis

After evaluating the OC of the 99 games in the study, we can also analyse the OC performance of all teams in the competition.

Fig. 4.4 compares the generated and received OC per effective time unit, for all teams. In the figure, the size of the points and their labels indicate the final ranking after the matches analysed. The dashed line plotted ($y = x$) separates the teams with positive and negative differences between the generated and received OC. Thus, teams below the dashed line generate less OC than their rivals, while teams above it accumulate more OC than their opponents.

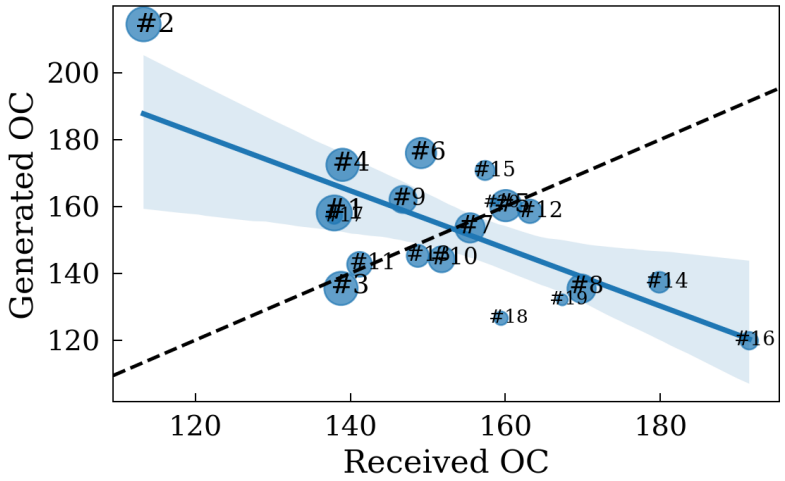


Fig. 4.4. Generated vs. received Offside Control. The values correspond to the average accumulated OC per effective time unit. The size of the nodes (and their labels) correlates with the ranking corresponding to all the matches analysed. The dashed line corresponds to $y = x$. The linear regression is shown in blue and has the equation $y = -0.86x + 285.42$, with $R^2 = 0.48$ and $RMSE = 14.96$. Teams above the dashed line have accumulated more Offside Controls than their rivals.

Ranking	Effective OC	Efficiency (%)
#1	160.23	84.35
#2	161.42	81.95
#3	137.31	87.75
#4	145.21	84.01
#5	157.14	82.97
#6	175.96	78.47
#7	172.44	81.67
#8	170.74	78.47
#9	132.05	84.32
#10	158.53	81.41
#11	142.67	78.57
#12	214.47	81.51
#13	126.55	84.49
#14	153.58	78.28
#15	135.48	81.46
#16	119.82	84.66
#17	135.42	81.62
#18	158.04	83.50
#19	162.09	82.75
#20	144.18	86.11

Table 4.1. AVERAGE EFFECTIVE OFFSIDE CONTROL VALUES AND EFFICIENCY FOR EACH TEAM FROM THE SPANISH NATIONAL LEAGUE (LALIGA SANTANDER) DURING THE SEASON 2019/2020, ORDERED BY THEIR RANKING

We can also observe a slight negative correlation between the OC generated and the OC received, although the correlation coefficient $R^2 = 0.48$ is relatively low. Note that the team in second place generates more OC and receives less. In addition, the teams at the top of the ranking are above the dashed line, with the

exception of the team in third place.

For a better understanding of the figure, Table 4.1 contains the Effective OC generated for each team, and the efficiency generating Offside Control, in descending order in ranking

One of the advantages of our proposed metric is its versatility: it can be used to describe the general behavior of the team, as presented above, but it can also be more specific and focus on the performance of specific players.

4.5. Performance measures using Offside Control

Fig. 4.5 shows the Effective Offside Control produced by two different strikers, identified as *A* and *B*. The left panel shows the positions on the pitch where the EOC is produced by each striker. The EOC is normalized by the total time played over the matches analyzed. In this way, we have a heat map showing where a striker tends to create controlled areas behind the offside line, which is useful information for analyzing his performance. In addition, the right panel shows the corresponding probability distribution functions of the time spent by both players at a given distance from the offside line. The distances are referenced to the offside line and are negative when the striker is in a valid position and positive when he is offside. Interestingly, striker *A* spends more time in front of the offside line than striker *B* (see the percentages of time spent on each side of the offside line). This means that player *A* is more likely to generate more EOCs.

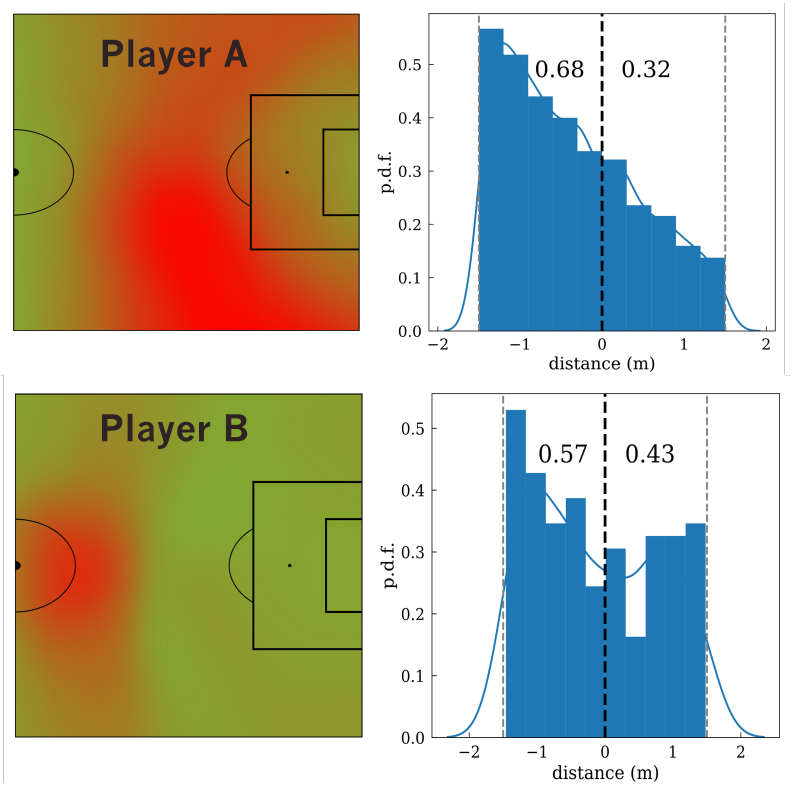


Fig. 4.5. Effective Offside Control (EOC) of two different strikers, A (upper plots) and B (lower plots). On the left, the positions on the pitch where the two players generate EOC. On the right, the probability distribution function of the time spent at a distance from the offside line. Negative values indicate that the player is in the correct position, while positive values indicate that the player is offside. The EOC is normalised to the time played. The values on the right plot indicate the proportion of time a player is in front of or behind the offside line.

We have been able to characterize the Offside Control of a total of 442 players. With all this data, we calculated the percentage of time that strikers spend close to the offside line and what percentage of that time the striker is in a valid position. We call this percentage a player's Offside Time Efficiency Ratio (OTER).

Next, we relate the OTER to the percentage of OC that is effective, i.e. EOC/IOC, and we call this percentage the *Offside Control Efficiency Ratio (OCER)*.

In Fig. 4.6 (A) we show the relationship between OTER and OCER for the 50 forwards[9]. We have obtained a positive correlation between the two metrics, as indicated by the solid blue line. This suggests that the more time a striker spends in a valid position near the offside line, the higher the probability of generating effective Offside Control.

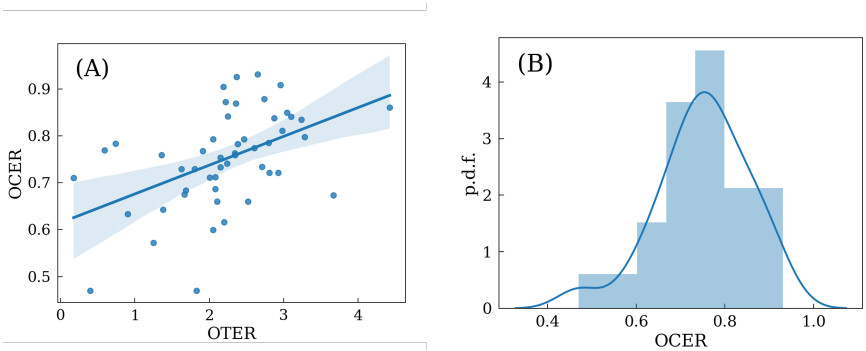


Fig. 4.6. Interplay between Offside Time Efficiency Ratio (OTER) and Offside Control Efficiency Ratio (OCER). On the left (A), we show the OCER vs. OTER of the strikers. On the right (B), the p.d.f of the OCER

However, the correlation coefficient between the two variables is low ($R^2 = 0.2325$, $p_{value} = 0.00039$), indicating that their relationship is complex and suggesting the effects of alternative variables not considered in our analysis. Finally, Fig. 4.6 (B) shows the probability distribution function of the values of the OCER. We can see that most of the Offside Control generated by the strikers is effective, indicated by a probability distribution mainly above 0.5 and a mean of $\langle OTER \rangle = 0.7506$. This result indicates that, in general, most of the pressure exerted by the strikers is effective.

[9]To ensure the statistical validity of the results, the study only included players who played at least 6 games.

5. CONCLUSION

At its core, Pitch Control model provides a quantitative framework using tracking data to assess player movement, spatial control, and decision making on the football field. Based on such a framework, we introduced the concept of *Offside Control* to analyze the performance of players with respect to the danger created behind the offside line of the opposing team.

The introduction of the OC parameter, derived from the modification of the Potential Pitch Control Field methodology, has provided a framework for studying the dynamics of pitch control beyond the offside line. By identifying attacking players within permissible boundaries and calculating the OC metric, this study has characterized the effectiveness of teams or individual players in exerting control beyond the offside line.

The results demonstrated the practical application of OC in various dimensions, from spatial distribution to temporal dynamics during matches. We also effectively characterized the distinction between effective and ineffective offside control, delineating between real surface control and ineffective control.

Furthermore, the versatility of OC was highlighted by its applicability at both the team and player levels. From tracking team-level OC performance over multiple matches to characterizing individual player behavior, OC emerged as a multifaceted metric capable of capturing diverse aspects of football.

The findings further underscored the potential utility of OC in assessing player efficiency and effectiveness near the offside line. The relationship between Offside Time Efficiency Ratio (OTER) and Offside Control Efficiency Ratio (OCER) provided insight into the interplay between player positioning and the generation of effective offside control. While a positive correlation between OTER and OCER was observed, the analysis also suggested the influence of additional variables, suggesting that the OC framework should be refined to capture such external influence.

Further work to develop this metric would include consideration of external factors, as well as improvements in tracking data collection and pitch control models. One of the major drawbacks of this methodology is that it cannot be performed in real-time during the game, due to the lack of real-time tracking data and the computational cost of calculating a Potential Pitch Control Surface.

We think that the development of both computer vision and the Pitch Control methodology will enable real-time analysis using tracking data, which is currently not possible. This will revolutionize data analysis in football, as all metrics and analyses previously defined in the literature for tracking data could be performed in real time, allowing for a broader understanding and control of football.

Overall, our study contributes to the ongoing discourse on tactical analysis in football by introducing a novel metric that enriches our understanding of the dynamics of the offside play. By revealing the intricacies of offside control and its implications for team and player performance, this research lays the groundwork for future investigations aimed at improving strategic insight and decision-making in football.

BIBLIOGRAPHY

Buldú, J. M., Busquets, J., Martínez, J. H., Herrera-Diestra, J. L., Echegoyen, I., Galeano, J., & Luque, J. (2018). Using network science to analyse football passing networks: Dynamics, space, time, and the multilayer nature of the game. *Frontiers in Psychology*, *9*. https://doi.org/10.3389/fpsyg.2018.01900

Felipe, J. L., García-Unanue, J., Viejo-Romero, D., Navandar, A., & Sánchez-Sánchez, J. (2019). Validation of a video-based performance analysis system (mediacoach®) to analyze the physical demands during matches in laliga. *Sensors (Basel, Switzerland)*, *19*. https://api.semanticscholar.org/CorpusID:202746198

Footballytics. (2021). *Data analytics in football*. https://www.footballytics.ch/post/data-analytics-in-football

Garrido, D., Burriel, B., Resta, R., del Campo, R. L., & Buldú, J. M. (2022). Heatmaps in soccer: Event vs tracking datasets. *Chaos, Solitons & Fractals*, *165*, 112827. https://doi.org/https://doi.org/10.1016/j.chaos.2022.112827

Novillo, Á., Gong, B., Martínez, J. H., Resta, R., del Campo, R. L., & Buldú, J. M. (2024). A multilayer network framework for soccer analysis. *Chaos, Solitons & Fractals*, *178*, 114355. https://doi.org/https://doi.org/10.1016/j.chaos.2023.114355

Pappalardo, L., Cintia, P., Rossi, A., Massucco, E., Ferragina, P., Pedreschi, D., & Giannotti, F. (2019). A public data set of spatio-temporal match events in soccer competitions. *Scientific Data*, *6*. https://doi.org/10.1038/s41597-019-0247-7

Sarlis, V., & Tjortjis, C. (2020). Sports analytics —evaluation of basketball players and team performance. *Information Systems*, *93*, 101562. https://doi.org/https://doi.org/10.1016/j.is.2020.101562

Spearman, W. (2018). Beyond expected goals.

Spearman, W., Basye, A., Dick, G., Hotovy, R., & Pop, P. (2017). Physics-based modeling of pass probabilities in soccer.

StatsPerform. (2023). *Opta data from stats perform*. https://www.statsperform.com/opta/

Wyscout. (2023). *Wyscout data glossary*. https://dataglossary.wyscout.com